Information Technology Laboratory

Information Access Division (IAD)

NIST
National Institute of Standards and Technology

Automatic Content Extraction 2004 Evaluation

Mark Przybocki and Audrey Le

The ACE Evaluation Workshop

Sept. 20-23, 2004

Hilton Alexandria Mark Center

ACE

The ACE program is dedicated to the development of technologies that automatically infer meaning from language data

ACE tasks

- There are four primary ACE Recognition tasks:
  - Entities
    - *Addressed in ACE-04*
    - *EDR, EMD, EDR co-reference*
  - Relations
    - *Addressed in ACE-04*
    - *RDR, RMD, RDR given ground truth entities*
  - Events
    - *Postponed until ACE-05, not yet satisfactorily defined*
  - Time Expressions
    - *Addressed in TERN-04*

The Recognition of ***Entities***

- The **E**ntity **D**etection and **R**ecognition task (**EDR**) measures a system’s ability to:
  - detect a **set of specified entities** mentioned in the source language,
  - recognize **selected information** about these entities. This information includes the *type*, *subtype*, *class* and *name(s)* of each entity, and also the entity mentions.

- The **E**ntity **M**ention **D**etection task (**EMD**) measures a system’s ability to:
  - correctly identify mentions of ACE entities

EDR – Entity Information

- TYPE [PER, ORG, LOC, GPE, FAC, VEH, WEA]
- SUBTYPE [*a different set for each* TYPE]
- CLASS [SPECIFIC *(others assigned a value of 0)*]
- {mentions}
  - TYPE [NAM, NOM, PRO, PRE]
  - ROLE [*Applied to GPE’s:* PER, LOC, ORG, GPE]
  - STYLE [LITERAL, METONYMIC]
  - head
  - extent [*entire nominal phrase*]
- {names}
  - name [*the proper name of a named entity*]

ACE ***Entity*** TYPES

PER	Humans either an individual or group
ORG	Groups defined by an organizational structure
VEH	Physical device primarily used to move an object
WEA	Physical device primarily used to harm / injure or destroy
GPE	Geographic regions defined by political and/or social groups
LOC	Geographic entities with physical extent
FAC	Permanent man-made structures

Ordered by decreasing evaluation value weights

*Full definitions of these entities may be found in the official annotation guidelines:
http://www.ldc.upenn.edu/Projects/ACE/Annotation/docs/

ACE *Entity* SUBTYPES

PER	(none)
ORG	Government, Commercial, Educational, Non-Profit, Other
VEH	Land, Air, Water, Subarea-Vehicle, Other
WEA	Blunt, Exploding, Sharp, Chemical, Biological, Shooting, Projectile, Nuclear, Other
GPE	Continent, Nation, Sate-or-Province, Count-or-District, Population-Center, Other
LOC	Address, Boundary, Celestial, Land-Region-Natural, region-Local, Region-Subnational, Region-National, Region-International, Water-Body, Other
FAC	Building, Subarea-Building, Bounded-Area, Conduit, path, Barrier, Plant, Other

ACE *Entity* CLASSES

- The CLASS describes the kind of reference the entity makes to something in the world
  - **Specific** *(value weight = 1.0)*
    - Refers to a particular or unique object
  - **Generic** *(value weight = 0)*
    - Refers to a kind or type of object
  - **Negative** *(value weight = 0)*
    - Refers to an empty set
  - **Under Specified** *(value weight = 0)*
    - Refers to an object that cannot be verified

ACE *Entity* Mentions

- Entity mention attributes
  - TYPE
    - The type of nominal phrase
    - NAM, NOM, PRE, PRO
  - ROLE
    - Applies to GPE’s
    - PER, LOC, ORG, GPE
  - STYLE
    - How the mention references the entity
    - LITERAL, METONYMIC
  - HEAD
    - The head of the nominal phrase
  - EXTENT
    - The entire nominal phrase

The Recognition of ***Relations***

- The <u>**R**elation **D**etection and **R**ecognition</u> task (**RDR**) measures a system’s ability to:
  - detect a **set of specified types of relations** mentioned in the source language,
  - recognize **selected information** about these relations. This information includes the *type, subtype* and *arguments* of each relation.
- The <u>**R**elation **M**ention **D**etection</u> task (**RMD**) measures a system’s ability to:
  - correctly identify mentions of ACE relations

ACE *Relation* TYPES and SUBTYPES

TYPE	SUBTYPE
PHYS *(Physical)*	Located, Near*, Part-whole
PER-SOC *(Personal / Social)*	Business*, Family*, Other*
EMP-ORG *(Employment/Membership/Subsidiary)*	Emp-Exec, Employ-Staff, Emp-Undet., Member-of-group, Partner*, Subsidiary, Other*
ART *(Agent-Artifact)*	User-or-Owner, Inventor-or-Manufacturer, Other
OTHER-AFF *(PER/ORG Affiliation)*	Ethnic, Ideology, Other
GPE-AFF *(GPE Affiliation)*	Citizen-or-Resident, Based-in, Other
DISC *(Discourse)*	(none)

* Denotes symmetric relations

*Full definitions of these relations may be found in the official annotation guidelines: http://www.ldc.upenn.edu/Projects/ACE/Annotation/

ACE – Input/Output

- Source Files
  - Three Languages: **English**, **Arabic**, and **Chinese**
  - Text documents (Broadcast News & Newswire)
    - English includes ASR version* of the Broadcast News data
  - UTF-8 Encoded
- Output Files
  - APF format that validates against the ACE DTD
  - Requires *Entity* and *Entity Mention* information
  - Optional *Relation* and *Relation Mention* information

* Thank-you to BBN for providing the ASR data estimated at 7-8% WER

ACE Evaluation

- Were all the reference entities correctly recognized?
  - A ***MISS**** occurs whenever a system misses an existing entity
    - one way that this can happen is for two distinct entities to be merged mistakenly into one
- Were all the system output entities valid entities?
  - A ***FALSE ALARM**** occurs whenever a system outputs an entity that doesn't exist
    - One way that this can happen is for one entity to be split mistakenly into two
- Were the valid system output entities correctly recognized?
  - An ***ERROR**** occurs whenever the TYPE, SUBTYPE or CLASS of the system entity doesn't match that of the reference entity.

* Similar terminology is used for "entity-mentions", "relations" and "relation mentions"

EDR Cost Model (1)

- The entity evaluation score is the sum of the values of ***all*** system ***output entities***

$$\text{EDR\_Value}_{sys} = \sum_{i} \text{value\_of\_sys\_entity}_i$$

- The overall score of a system is computed as the system output information relative to perfect output:

$$\text{System\_Value}= \frac{\sum_{i} \text{value}(\text{sys\_output}_i, \text{reference}_{map(i)})}{\sum_{m} \text{value}(\text{reference}_m, \text{reference}_m)}$$

EDR Cost Model (2)

- The value of ***each*** system ***output entity*** is the product of an inherent ***entity value*** and the sum of the values of the ***entity's mentions***

$$\text{Value}_{\text{sys\_entity}} = \text{Entity\_Value(sys\_entity)} * \sum_{m} \text{Mention\_Value}(\text{sys\_mention}_m)$$

EDR Cost Model (3)

- The ***entity_value*** of a system output entity is a function of its type
  - If the output entity is mapped, then the minimum value for the system entity and its corresponding reference entity is used (discounted if errors in *type*, *subtype* and *class*)
  - If unmapped, it is weighted by a false alarm penalty

$$\text{ENT\_VAL} = \begin{cases} \min\left(\begin{array}{l} \text{E\_TypeVal(sys)} * \\ \text{E\_ClassVal(sys)}, \\ \text{E\_TypeVal}(\text{ref}_{\text{sys}}) * \\ \text{E\_ClassVal}(\text{ref}_{\text{sys}}) \end{array}\right) * (W_{\text{E-err-type}} * W_{\text{E-err-subtype}} * W_{\text{E-err-class}}) & \textit{(when mapped)} \\ \text{E\_TypeVal(sys)} * \text{E\_ClassVal(sys)} * W_{\text{E-FA}} & \textit{(when not mapped)} \end{cases}$$

EDR Cost Model (4)

- The ***mention_value*** of a system entity mention is a function of its type
  - If the mention is mapped, then the minimum value for the sys mention and its corresponding ref mention is used
    - Mention_Value is discounted for errors in mention *type*, *role* and *style*
  - If unmapped, it is weighted by a false alarm penalty

$$\text{MEN\_VAL} = \begin{cases} \min\begin{pmatrix} \text{M\_TypeVal(sys)}, \\ \text{M\_TypeVal}(\text{ref}_{\text{sys}}) \end{pmatrix} * (W_{\text{M-err-type}} * W_{\text{M-err-role}} * W_{\text{M-err-style}}) & \textit{(when mapped)} \\ \text{M\_TypeVal(sys)} * (W_{\text{M-FA}} * W_{\text{M-CoRef}}) & \textit{(when not mapped)} \end{cases}$$

RDR Cost Model (1)

- The relation evaluation score is the sum of the values of ***all*** system ***output relations***

$$\mathrm{RDR\_Value}_{sys} = \sum_{i} \mathrm{value\_of\_sys\_relation}_i$$

- The overall score of a system is computed as the system output information relative to perfect output:

$$\mathrm{System\_Value} = \frac{\sum \mathrm{value}(\mathrm{sys\_output}_i, \mathrm{reference}_{map(i)})}{\sum \mathrm{value}(\mathrm{reference}_m, \mathrm{reference}_m)}$$

RDR Cost Model (2)

- The value of ***each*** system ***output relation*** is the product of an inherent ***relation value*** and the sum of the values of the ***relation’s entity arguments***

$\text{Value}_{\text{sys\_relation}} =$	$(\text{Relation\_Value}(\text{sys\_relation})) *$ $(\sum_{a} \text{Argument\_Value}(\text{sys\_argument}_{a}))$

RDR Cost Model (3)

- The ***relation_value*** of a system output relation is a function of its type
  - If the output relation is mapped, then the minimum value for the system relation and its corresponding reference relation is used (discounted if errors in *type* and *subtype*)
  - If unmapped, it is weighted by a false alarm penalty

$$\text{REL\_VAL} = \begin{cases} \min\begin{pmatrix} \text{R\_TypeVal(sys)}, \\ \text{R\_TypeVal}(\text{ref}_{\text{sys}}) \end{pmatrix} * (W_{\text{R-err-type}} * W_{\text{R-err-subtype}}) & \textit{(when mapped)} \\ \text{R\_TypeVal(sys)} * W_{\text{R-FA}} & \textit{(when not mapped)} \end{cases}$$

RDR Cost Model (4)

- The ***argument_value*** of a system relation argument is the ***entity_value*** of that entity argument, where the entity argument of the system relation is mapped to the corresponding argument of the reference relation

Argument_Value = Entity_Value(sys)

Mapping System Output to Reference

- System ***entities*** are mapped to reference ***entities*** so as to maximize **EDR** value
- System ***relations*** are mapped to reference ***relations*** so as to maximize **RDR** value

ACE-EVAL Parameter Settings (1)

Entity Mention	value weight
NAM	1.000
NOM	0.200
PRE	0.200
PRO	0.040
Entity Types	value weight
PER	1.000
ORG	0.500
VEH	0.500
WEA	0.500
GPE	0.250
LOC	0.100
FAC	0.050

Entity Classes	value weight
SPECIFIC	1.000
all others	0.000
Entity Attribute	value discount
CLASS	0.750
SUBTYPE	0.900
TYPE	0.500
Mention Attribute	value discount
ROLE	0.900
STYLE	0.900
TYPE	0.900
Other Costs	
False Alarm Entity	0.750
False Alarm Mention	0.750
Discount Incorrect coref	0.000

ACE-EVAL Parameter Settings (2)

Relation Types	value weight
ART	1.000
DISC	1.000
EMP-ORG	1.000
GPE-AFF	1.000
METONYMY	1.000
OTHER-AFF	1.000
PER-SOC	1.000
PHYS	1.000

Relation Attribute	value discount
SUBTYPE	0.900
TYPE	0.500
Other Costs	
False Alarm Relation	0.750

ACE Tools

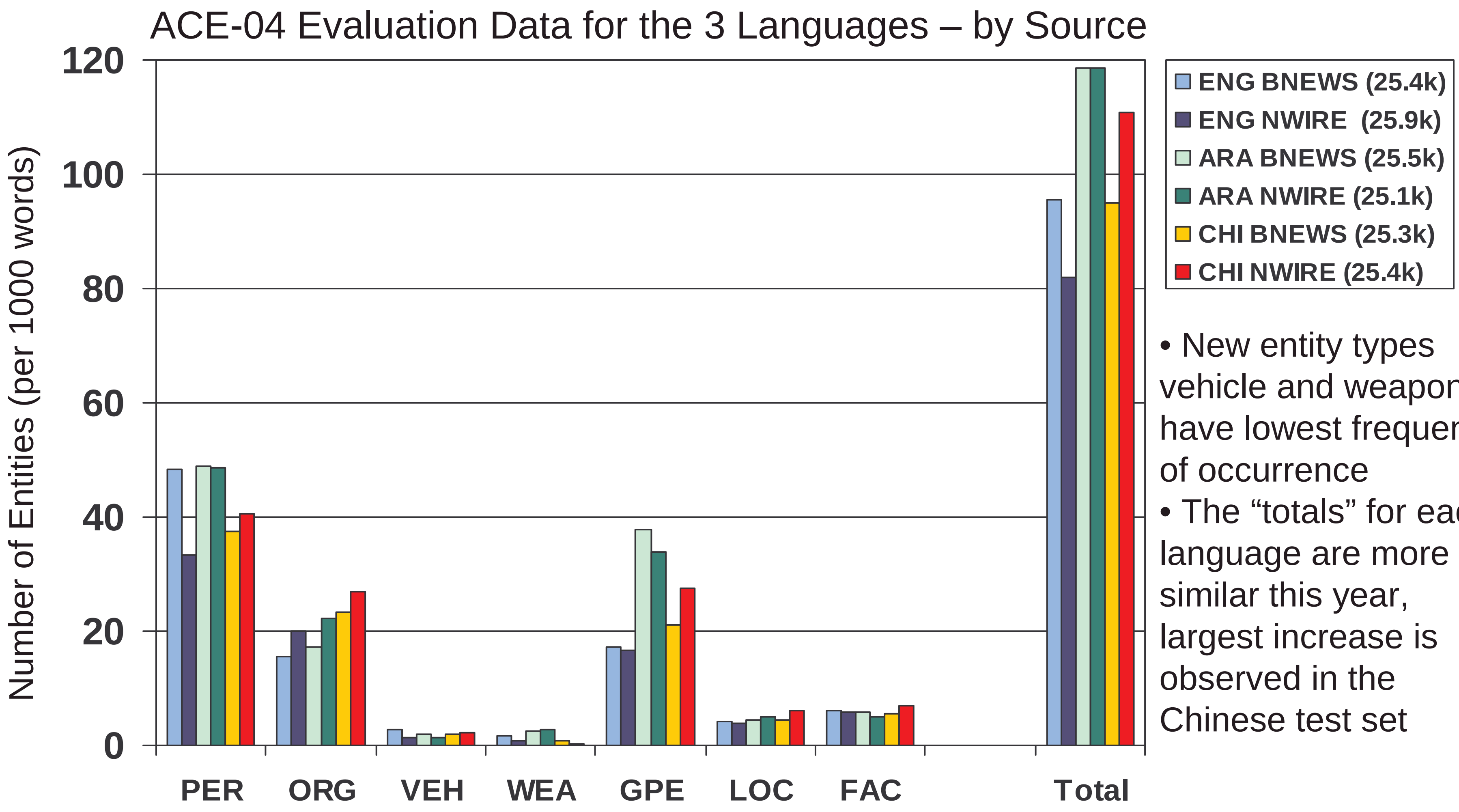
- ace-eval-v10.pl
  - Official scoring script used for ACE-04
- apf-v4.0.1.dtd
  - Current ACE DTD
- xmlvalid
  - A java based XML validation program which is used to validate ACE hypothesis and reference files
  - To be distributed with future test sets so participants can be sure they are submitting valid ACE APF files

ACE Data – Research Corpora

Training Data (Oct–Dec 2000)	
English Resources	
BNews	60,291 words
Newswire	59,840 words
Treebank translations Fisher conversations	37,822 words
Arabic Resources	
BNews	63,238 words
Newswire	63,122 words
Treebank	25,010 words
Chinese Resources	
BNews	~67,702 words
Newswire	~60,251 words
Treebank	~25,749 words

Evaluation Data (Jan 2001)	
English Resources	
BNews	25,365 words
Newswire	25,926 words
STT of BNews	~25,000 words
Arabic Resources	
BNews	25,471 words
Newswire	25,056 words
Chinese Resources	
BNews	25,318 words
Newswire	25,379 words

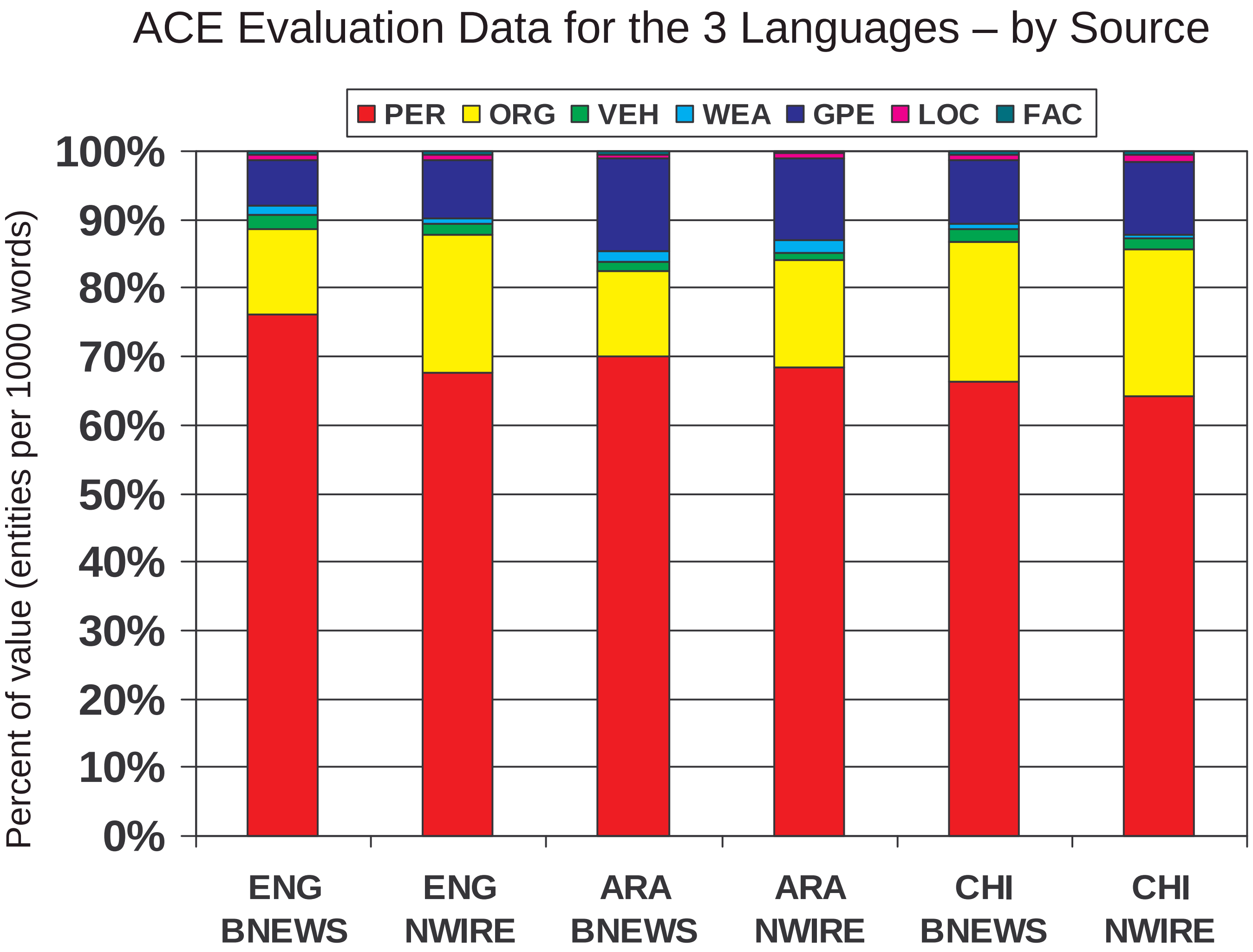
Distribution of ***Entity Types***



- New entity types vehicle and weapon have lowest frequency of occurrence
- The "totals" for each language are more similar this year, largest increase is observed in the Chinese test set

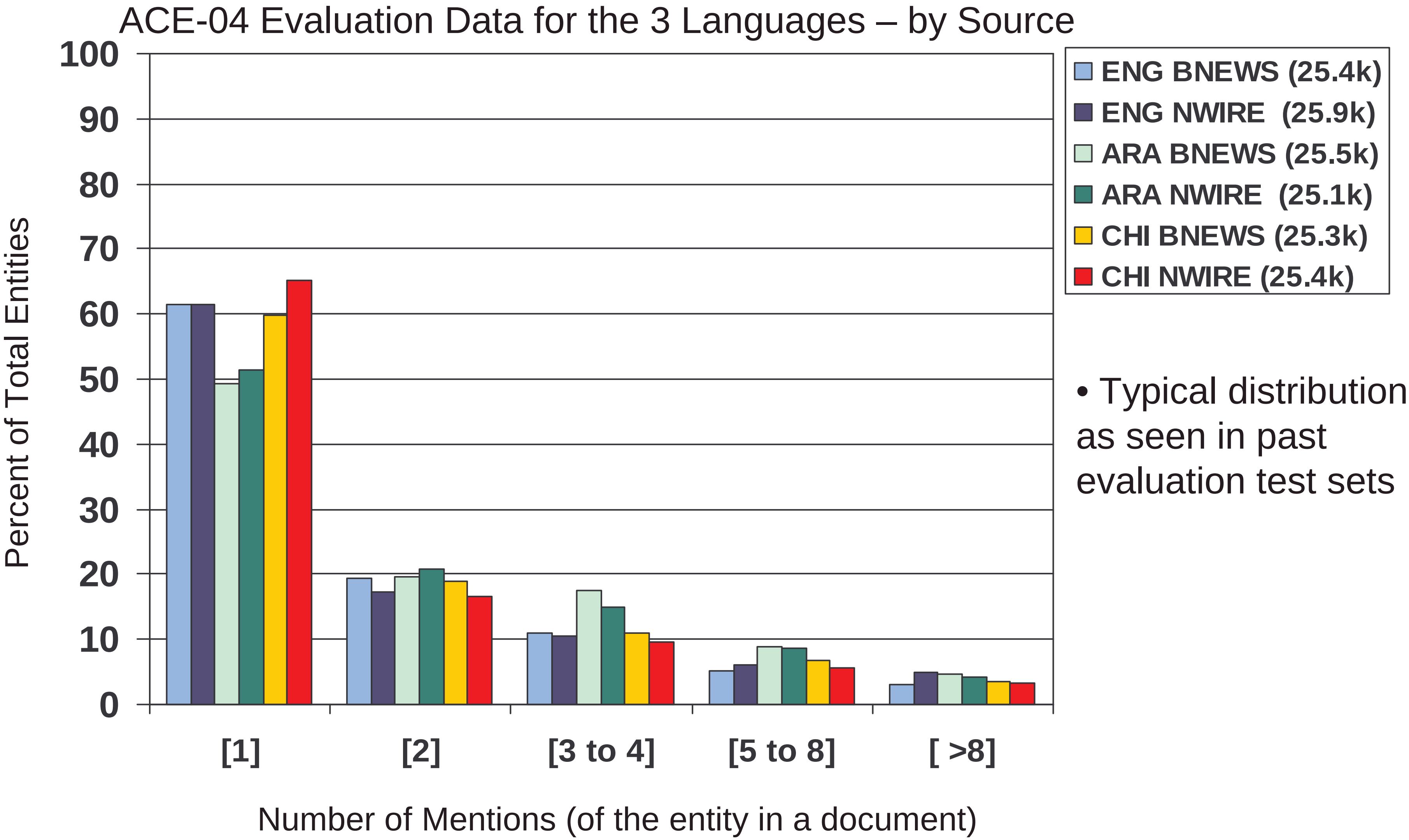
Entity Types, ordered by decreasing evaluation importance (scoring weights)

Distribution of ***Entity Types***



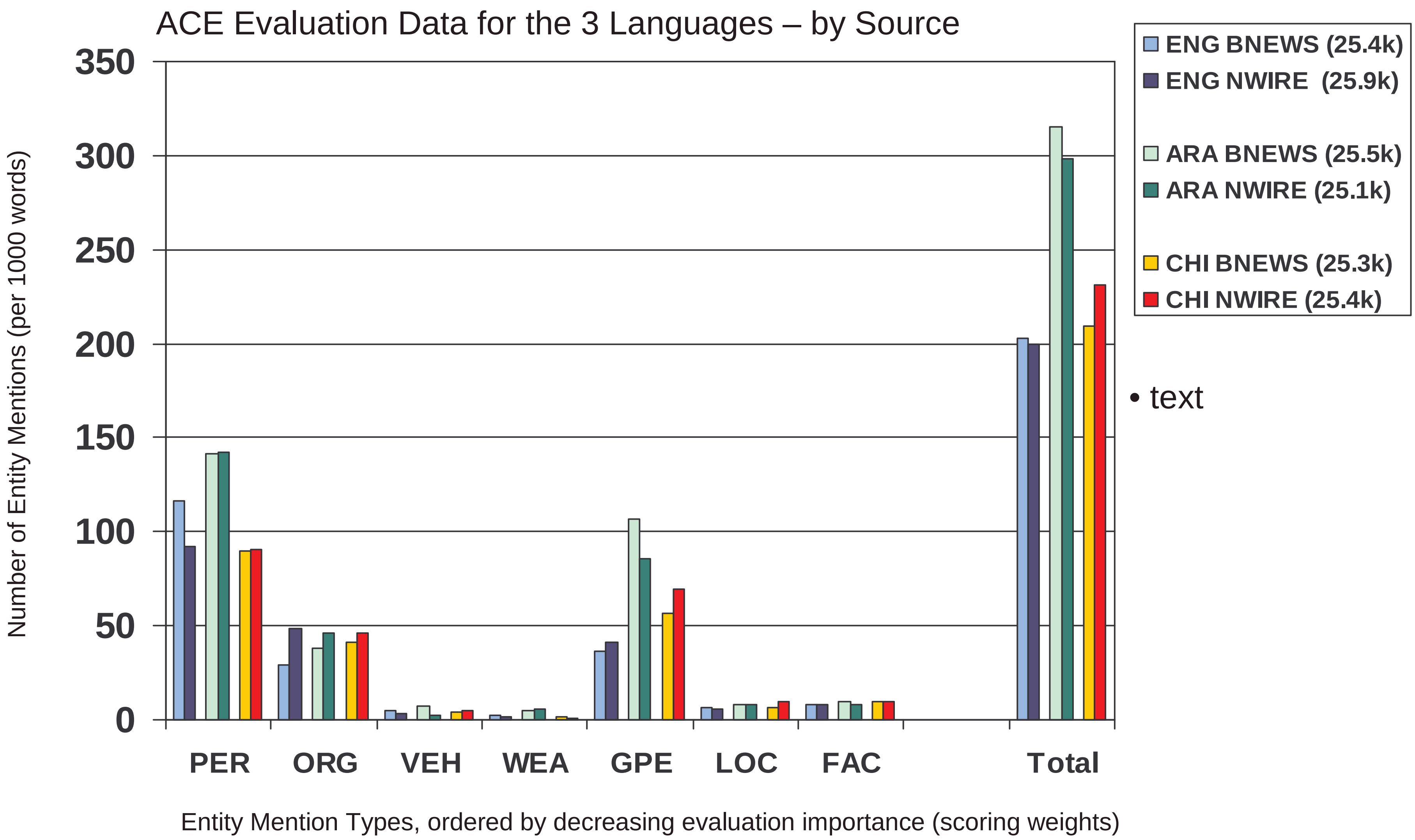
- Distribution is in terms of percent of value
- PER entities dominate the overall value, followed by ORG and GPE
  - Other types do not contribute significantly to value

Distribution of ***Entities*** by Mention Count



- Typical distribution as seen in past evaluation test sets

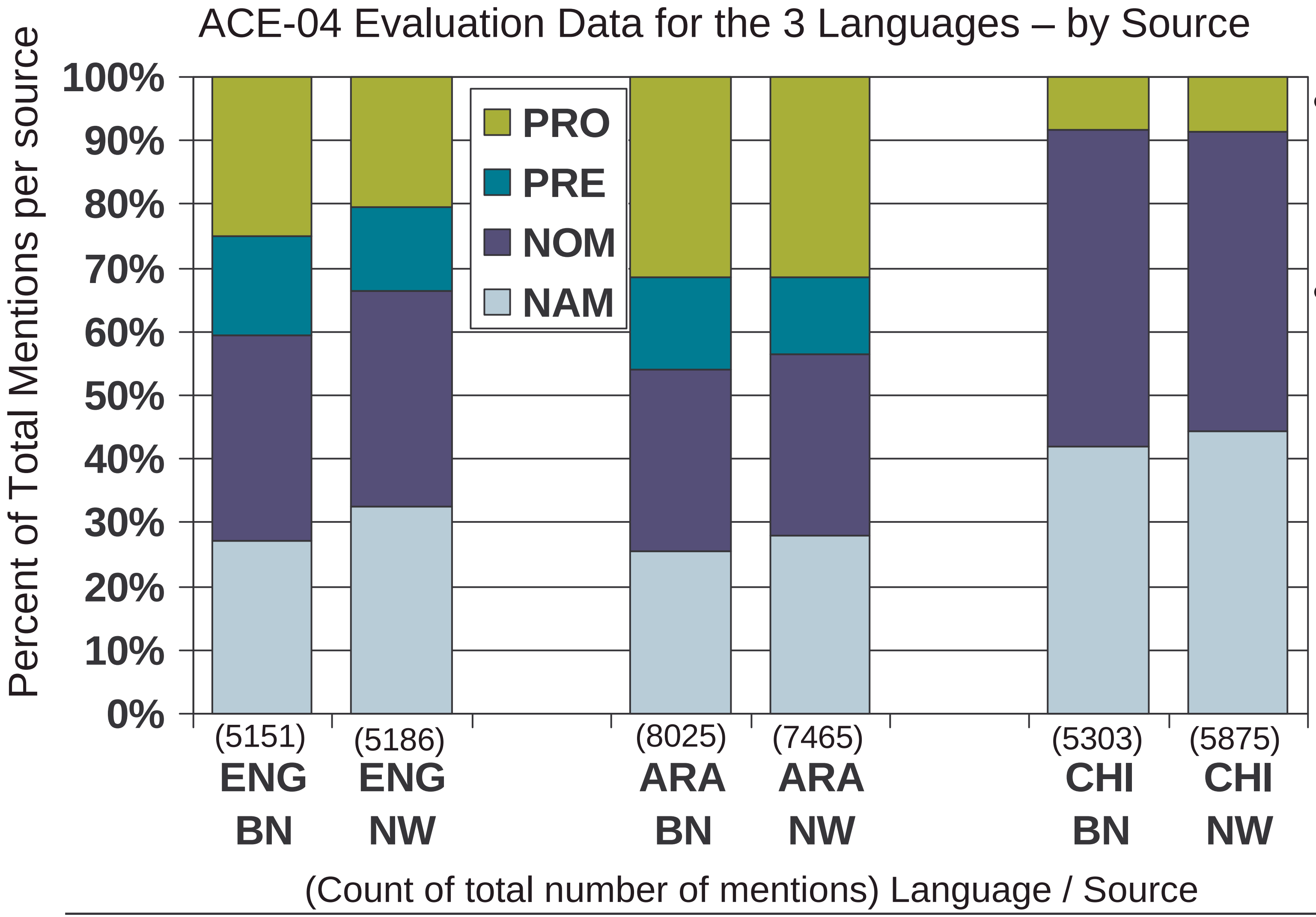
Distribution of ***Entity Mention*** Types



ACE Evaluation Data for the 3 Languages – by Source

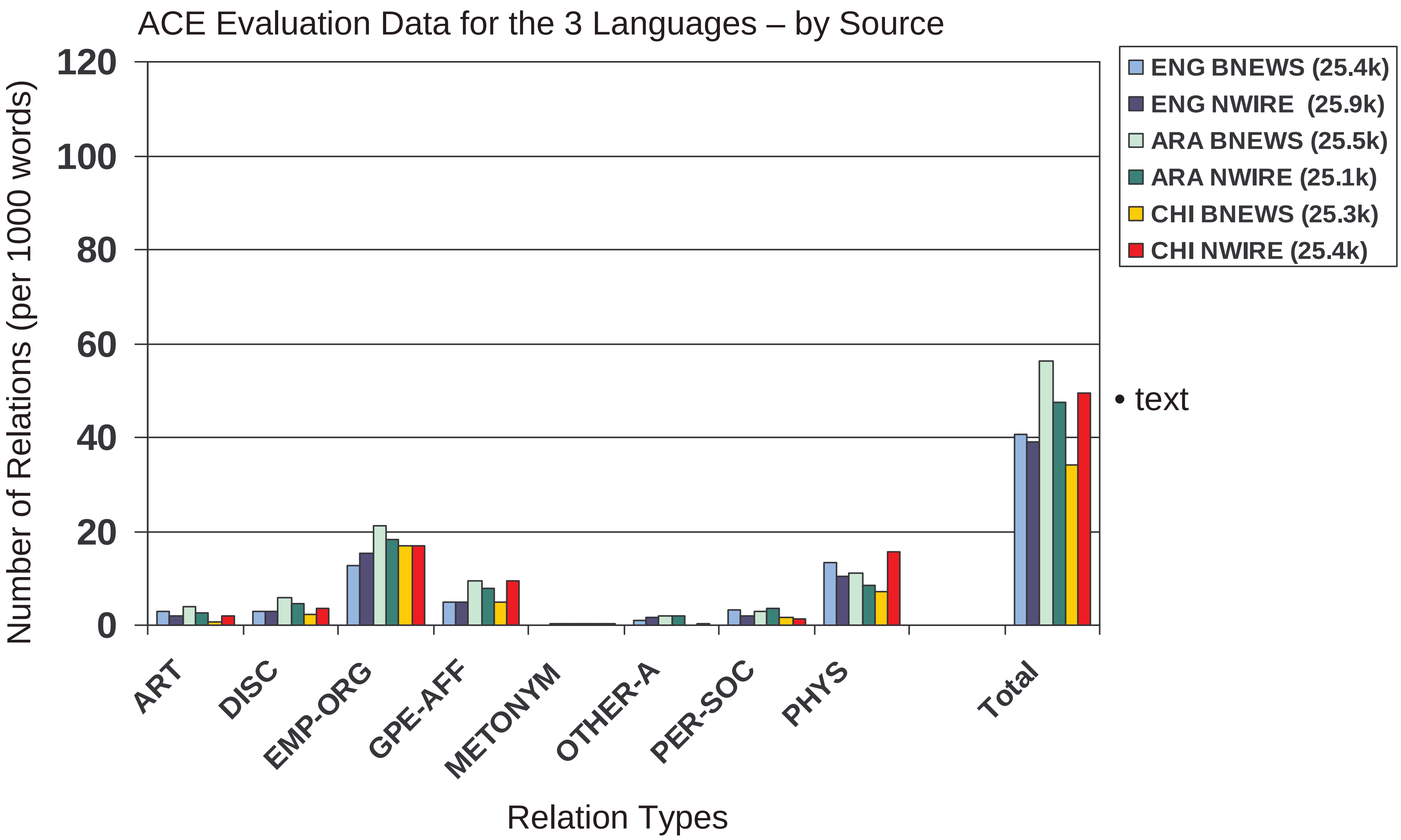
- text

Distribution of ***Mention Levels***



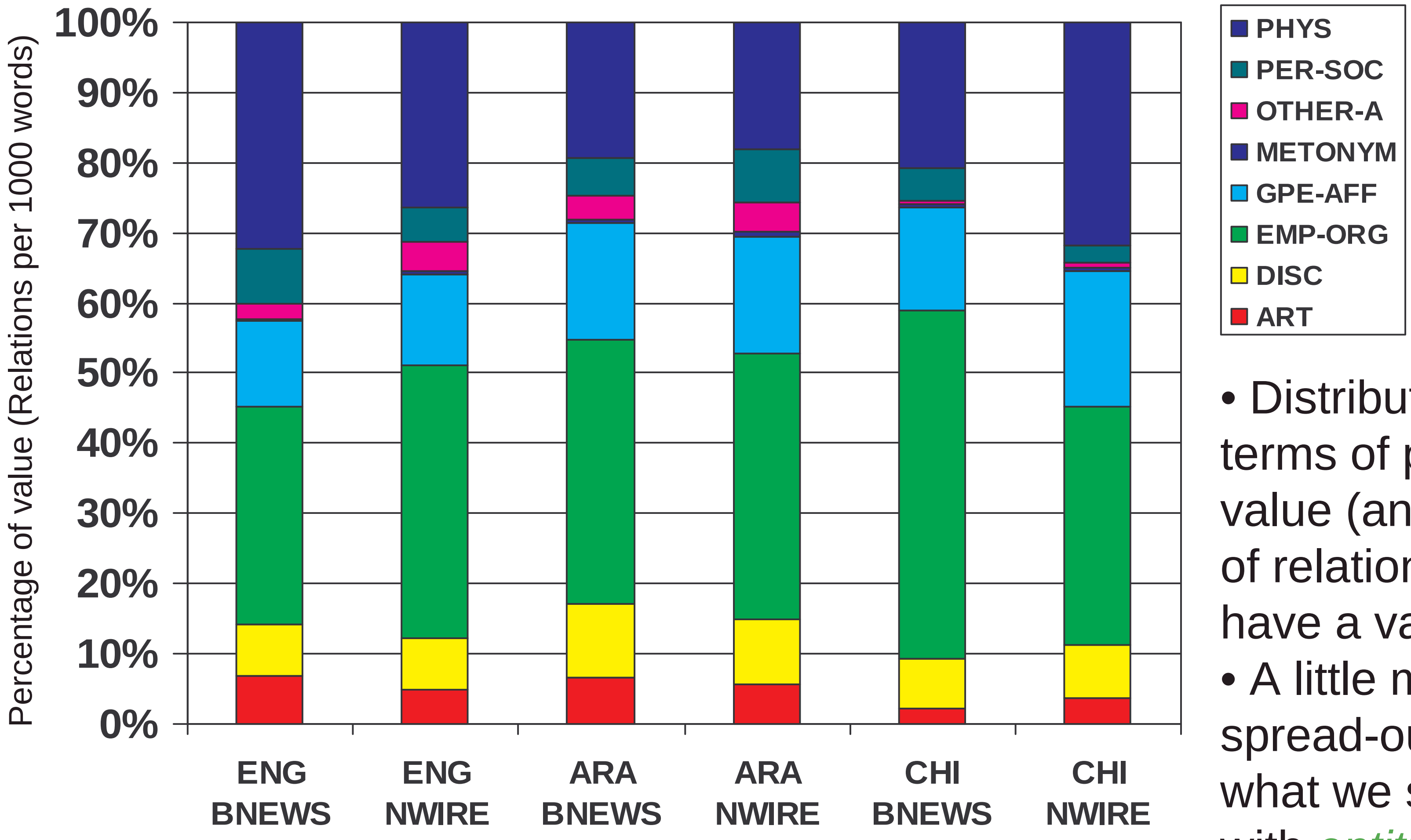
- More mentions for Arabic
- No PRE mentions for Chinese

Distribution of *Relation Types*



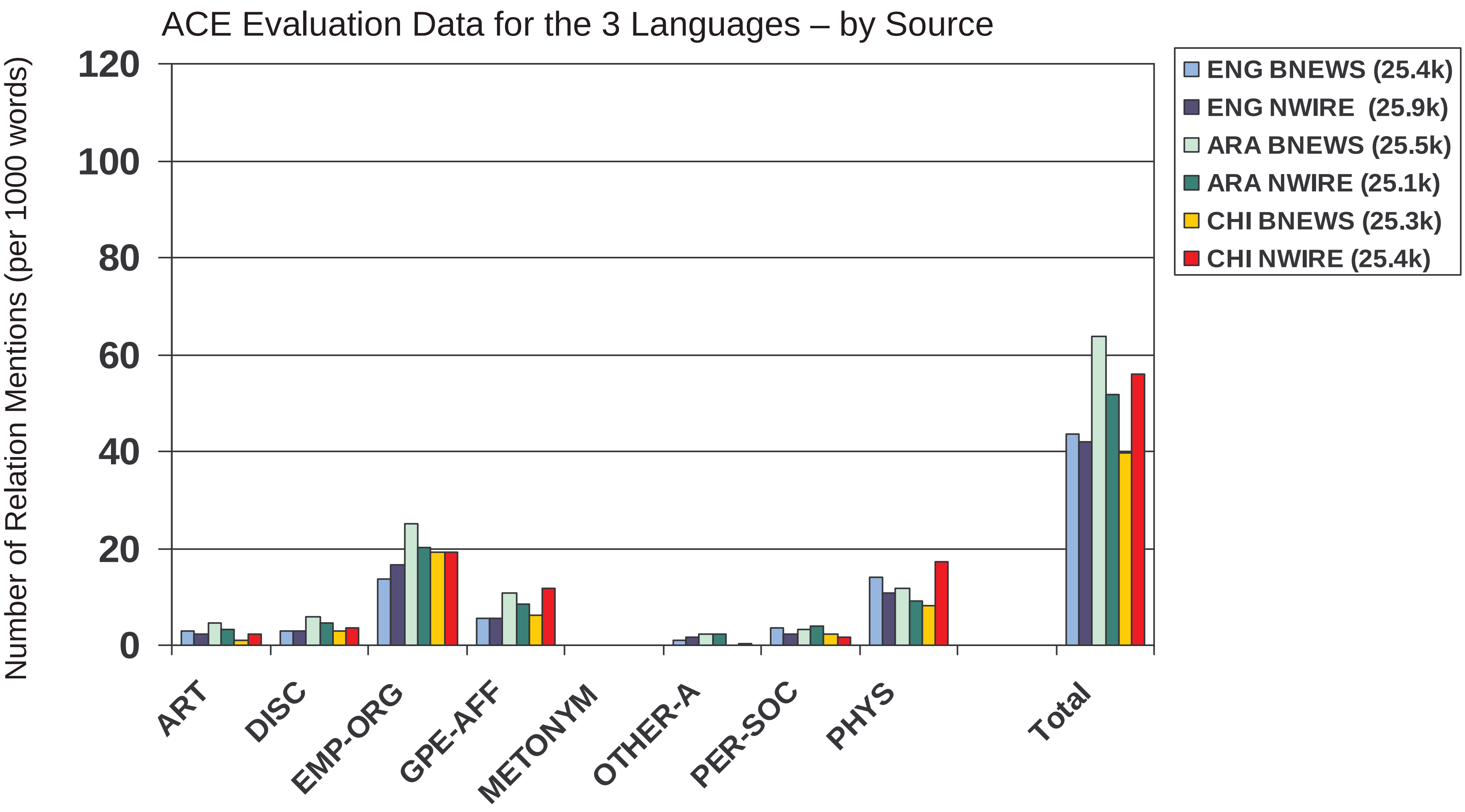
- text

Distribution of *Relation Types*



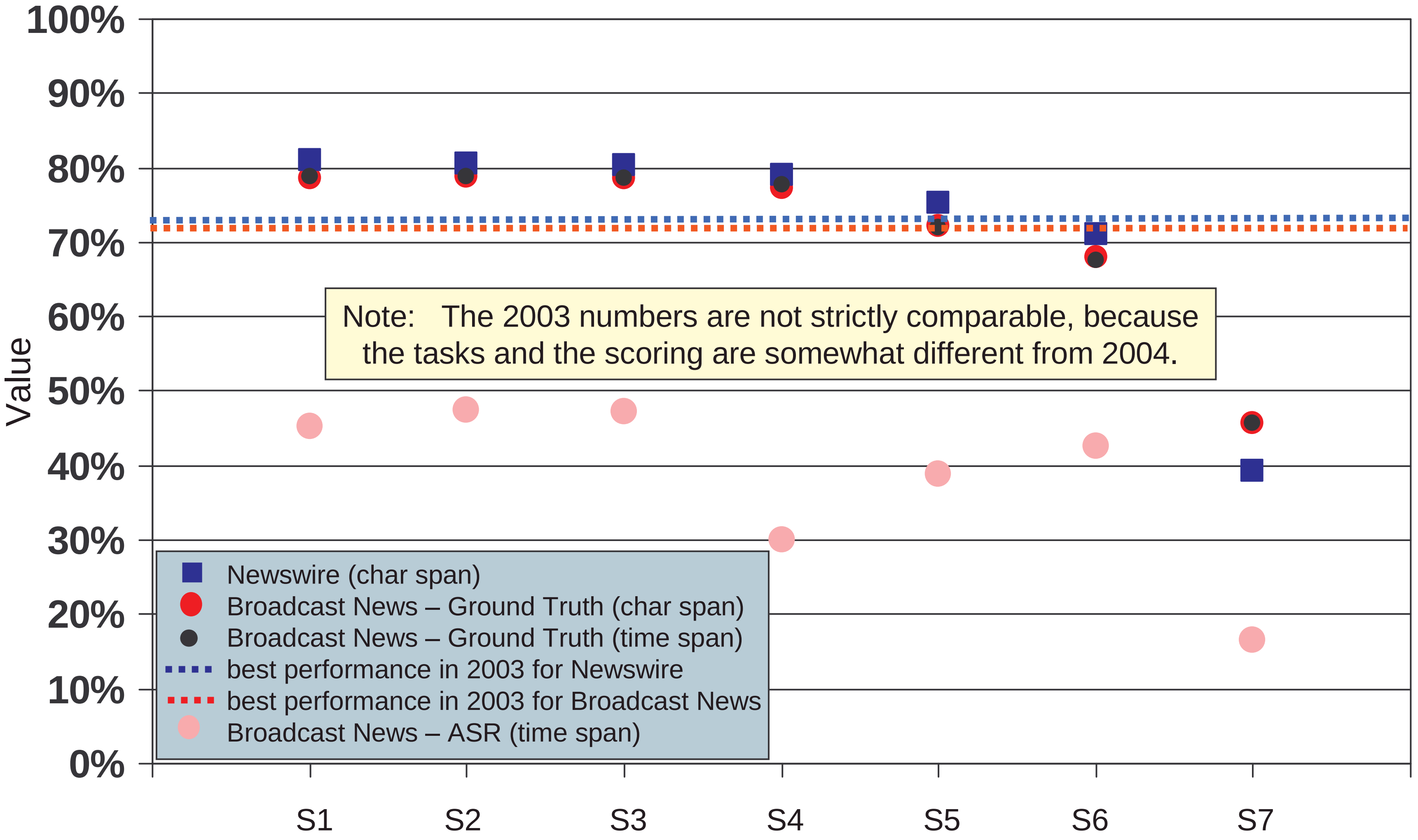
- Distribution in terms of percent of value (and number of relations since all have a value of 1.0)
- A little more spread-out then what we saw with *entities*

Distribution of ***Relation Mention*** Types



Relation Mention Types, ordered by decreasing evaluation importance (scoring weights)

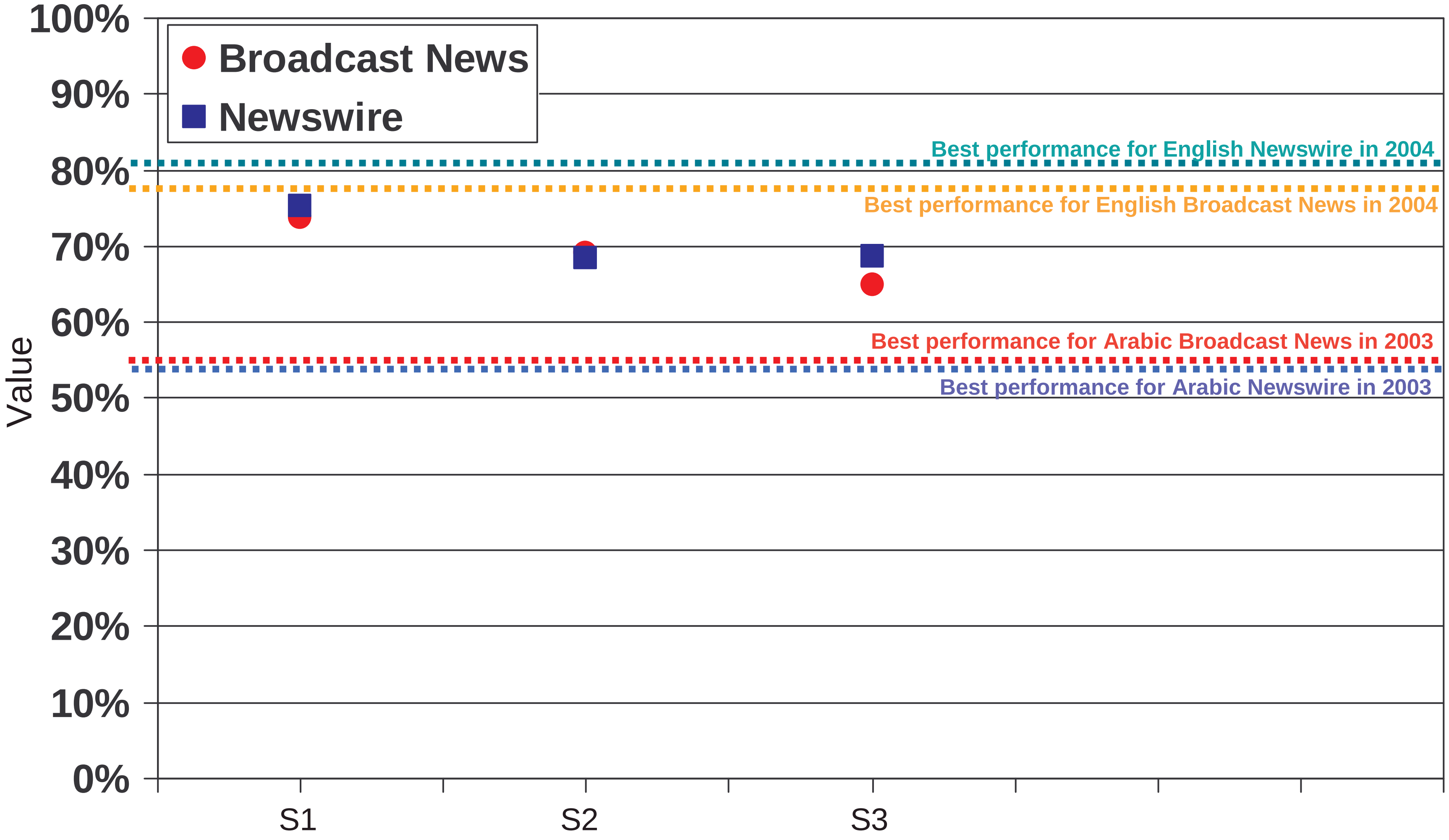
EDR Results for *English*



Note: The 2003 numbers are not strictly comparable, because the tasks and the scoring are somewhat different from 2004.

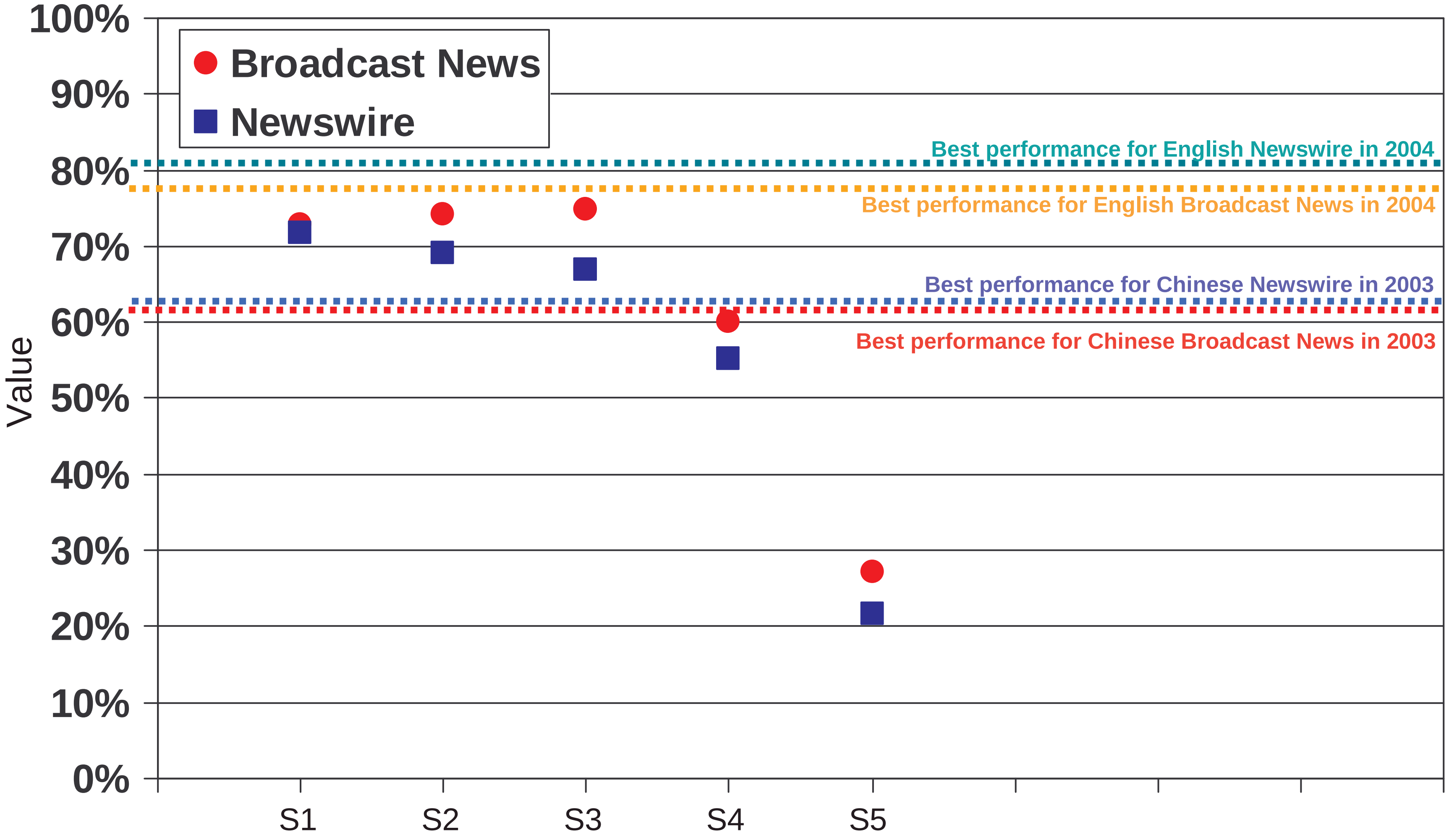
2004 EDR Results English – Source Set - systems ordered by decreasing average % value (combined)

EDR Results for *Arabic*



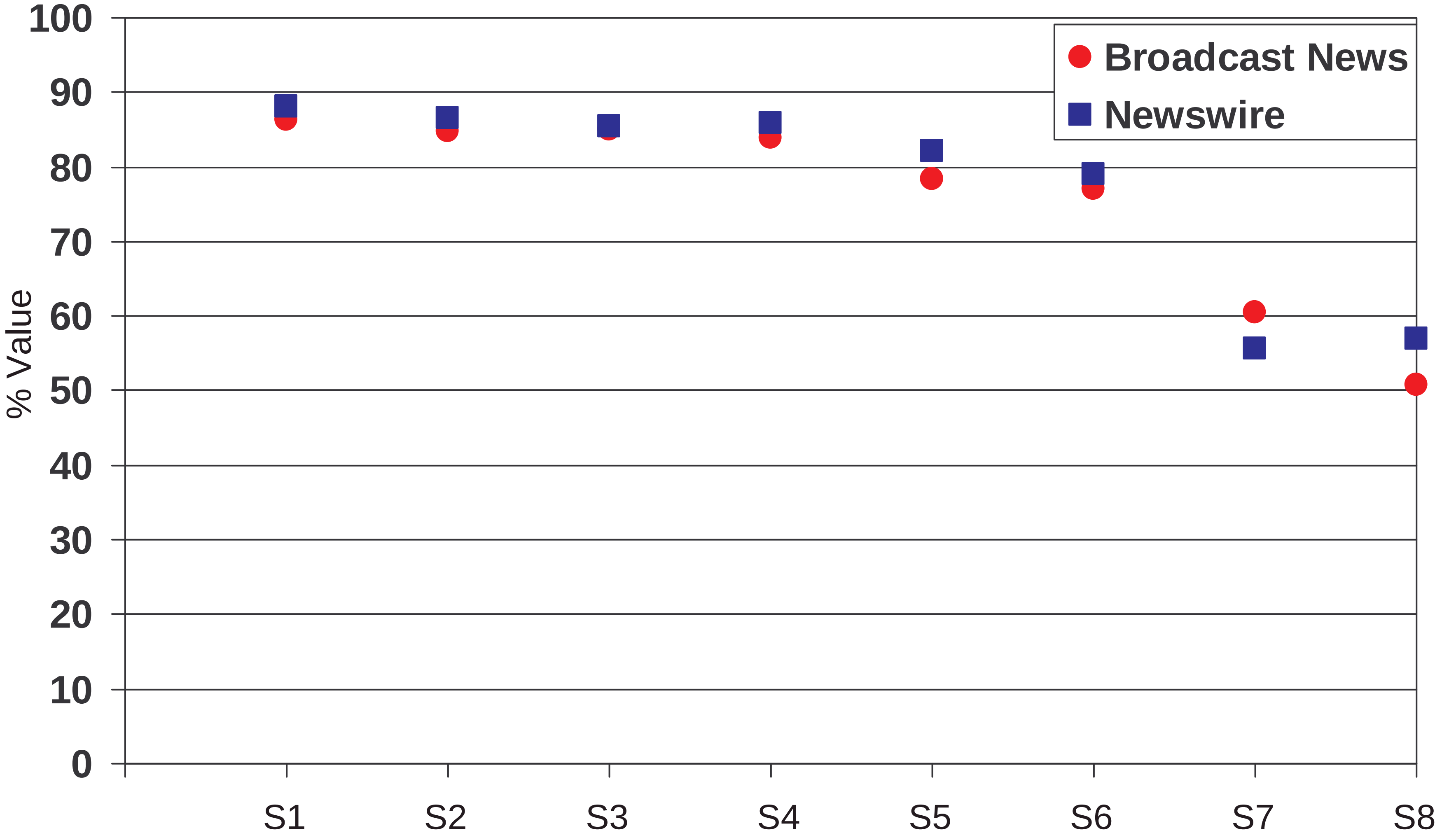
2004 EDR Results Arabic – Source Set - systems ordered by decreasing average % value (combined)

EDR Results for *Chinese*



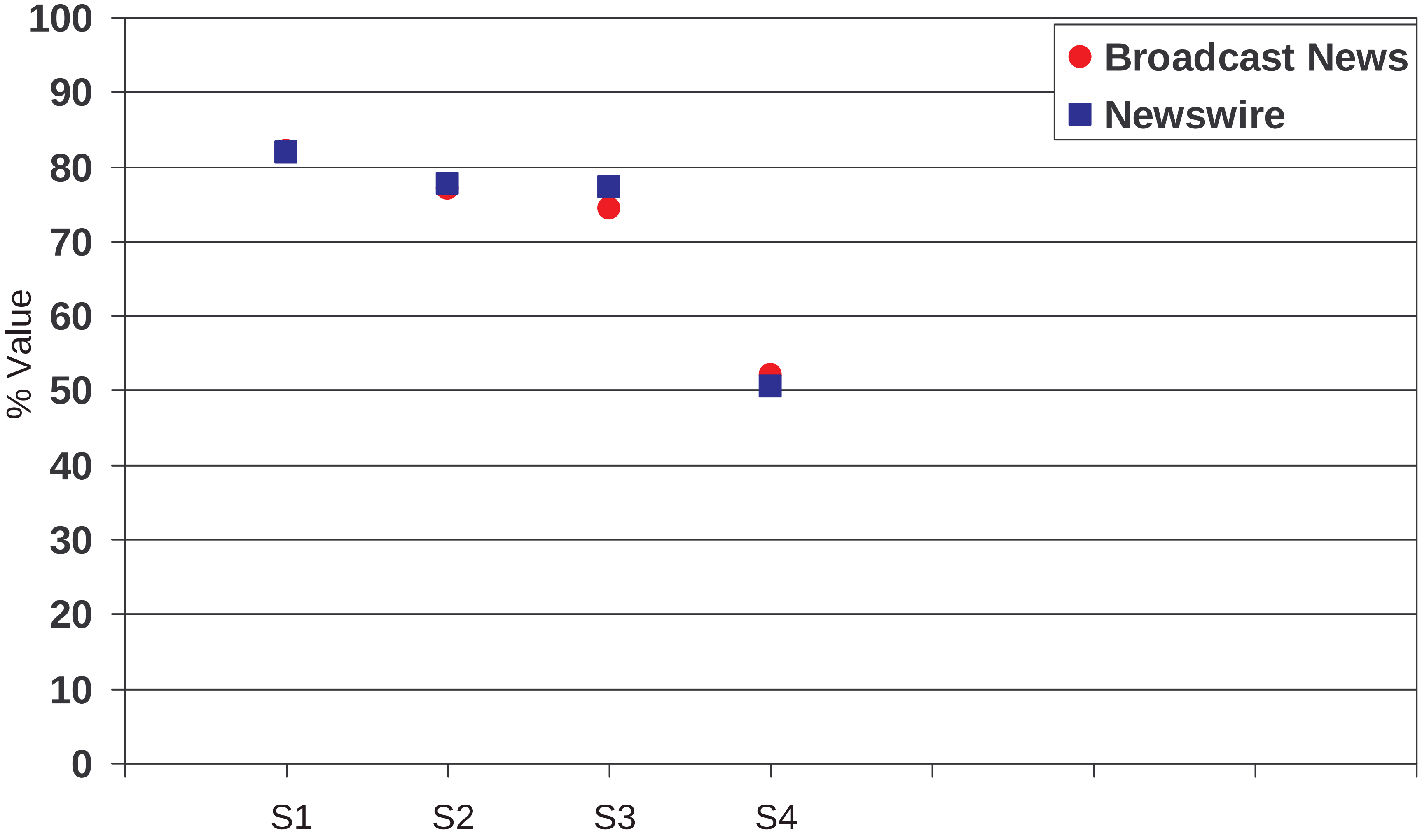
2004 EDR Results Chinese – Source Set - systems ordered by decreasing average % value (combined)

EMD Results for *English*



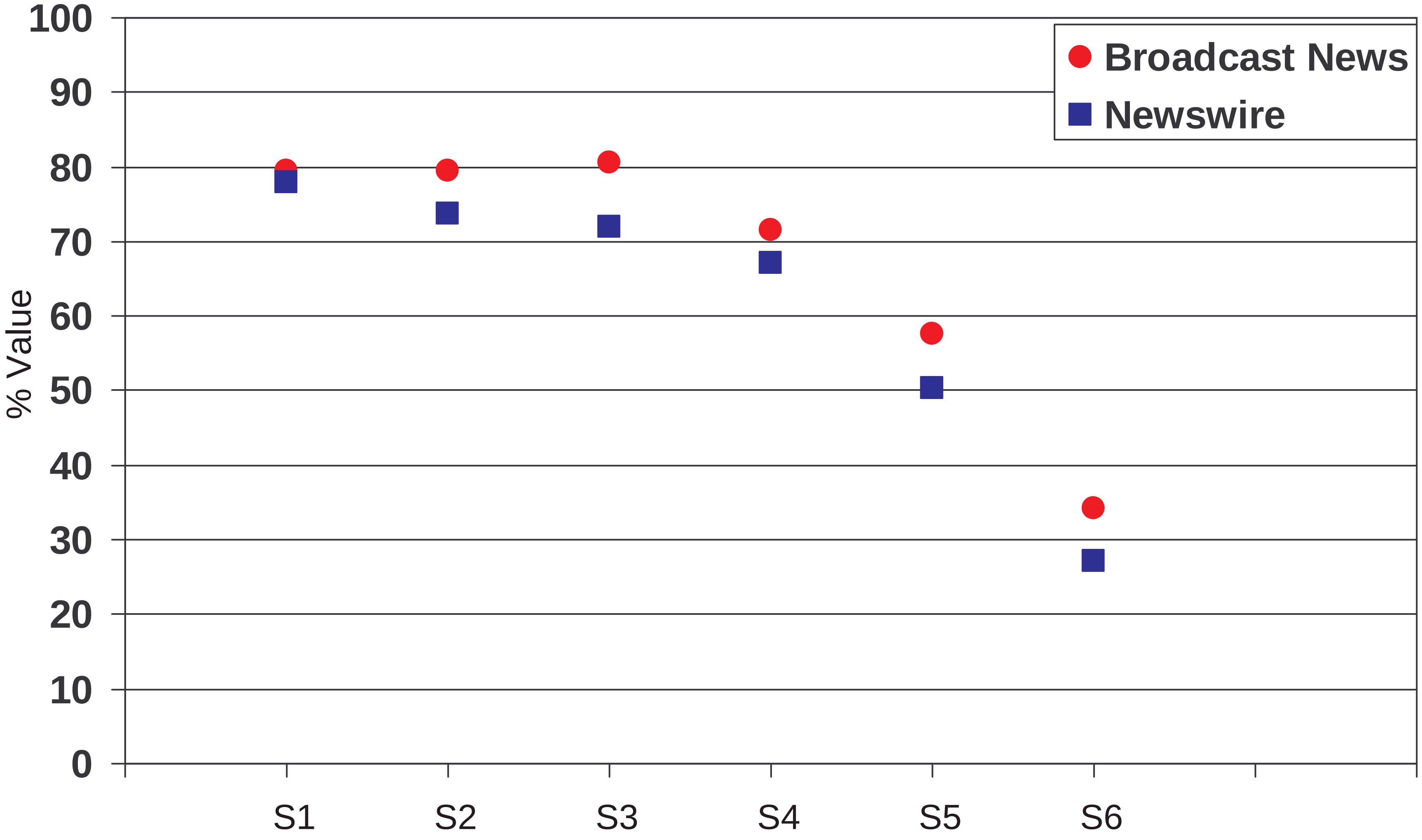
2004 EMD Results English – Source Set - systems ordered by decreasing average % value (combined)

EMD Results for *Arabic*



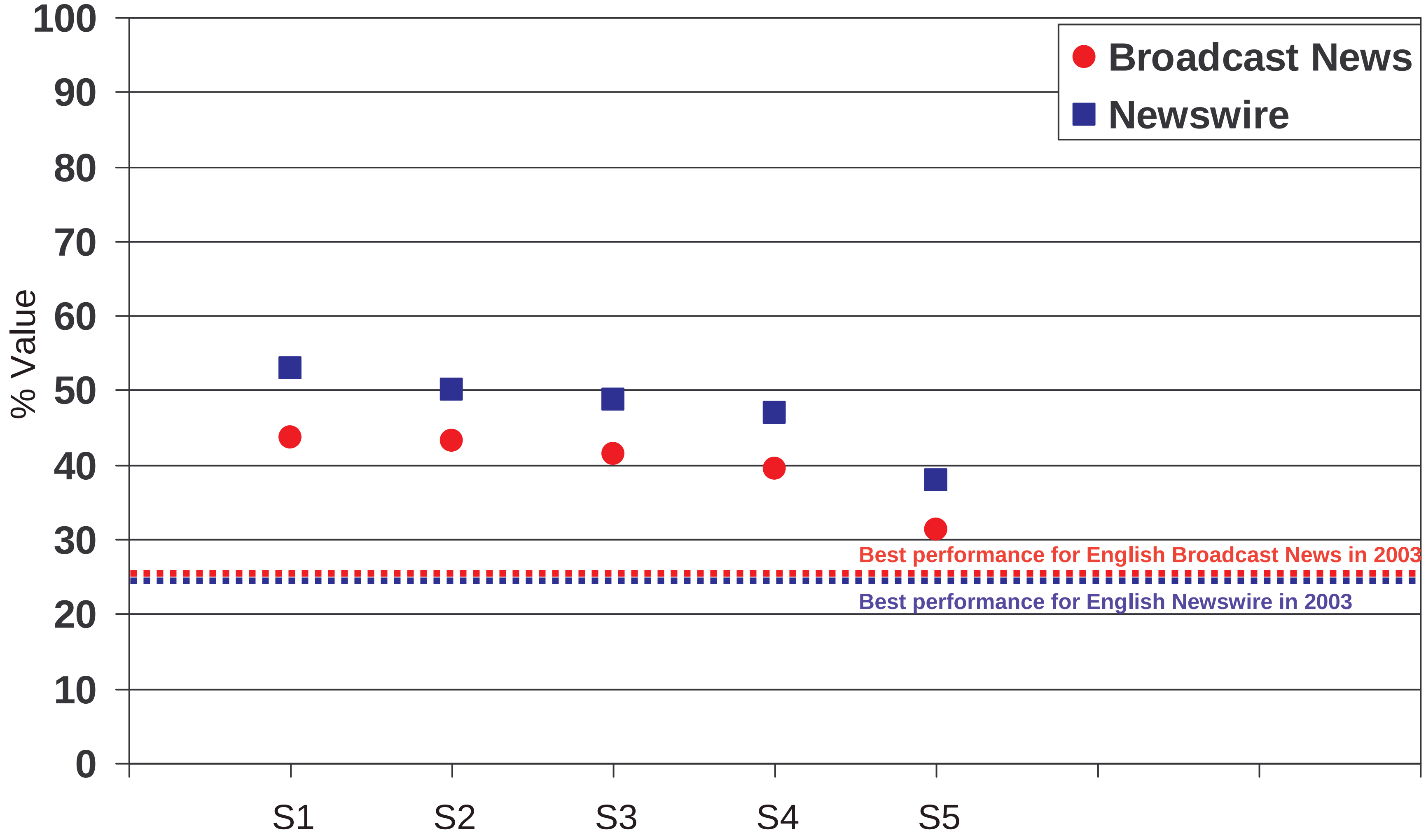
2004 EMD Results Arabic – Source Set - systems ordered by decreasing average % value (combined)

EMD Results for *Chinese*



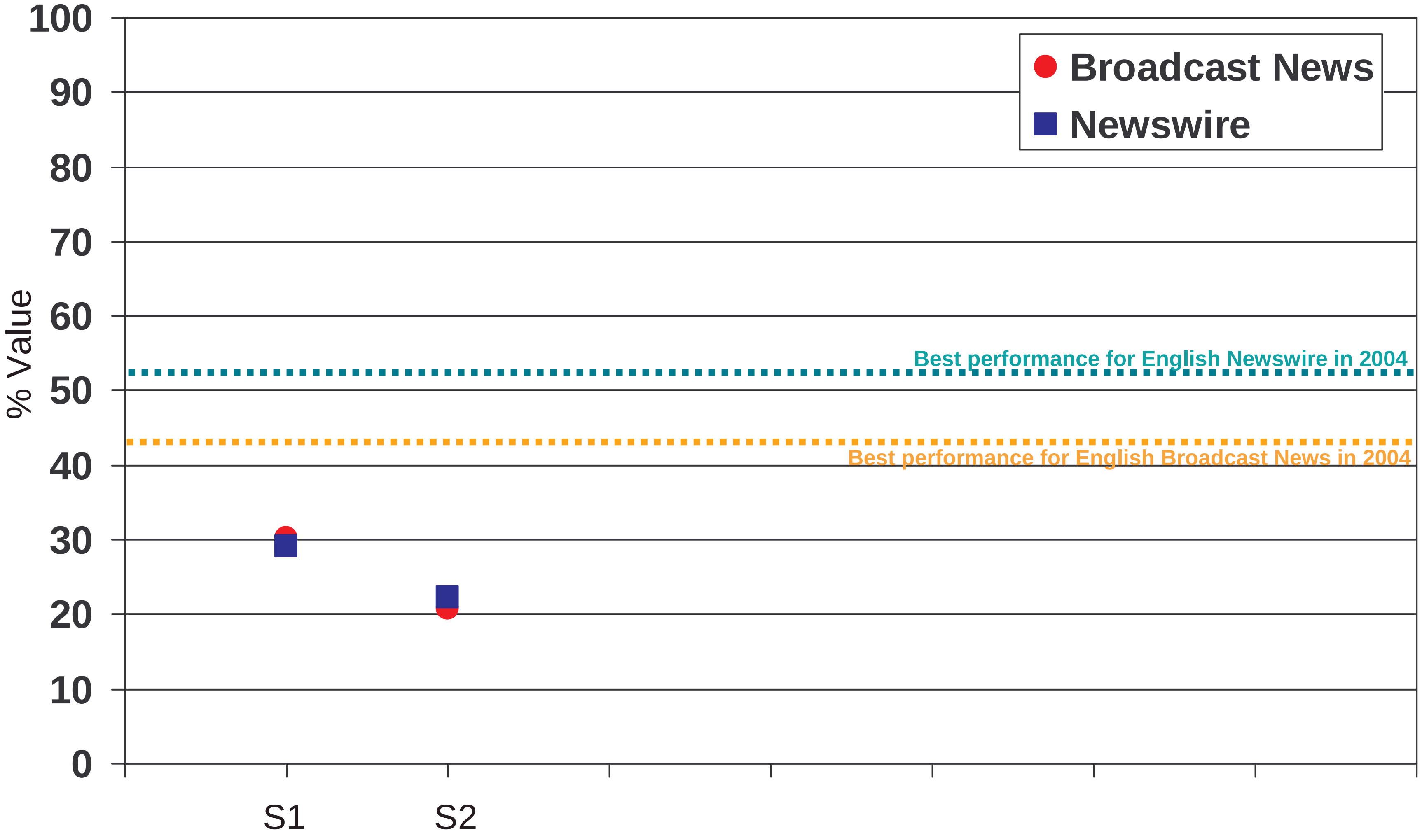
2004 EMD Results Chinese – Source Set - systems ordered by decreasing average % value (combined)

RDR Results for *English*



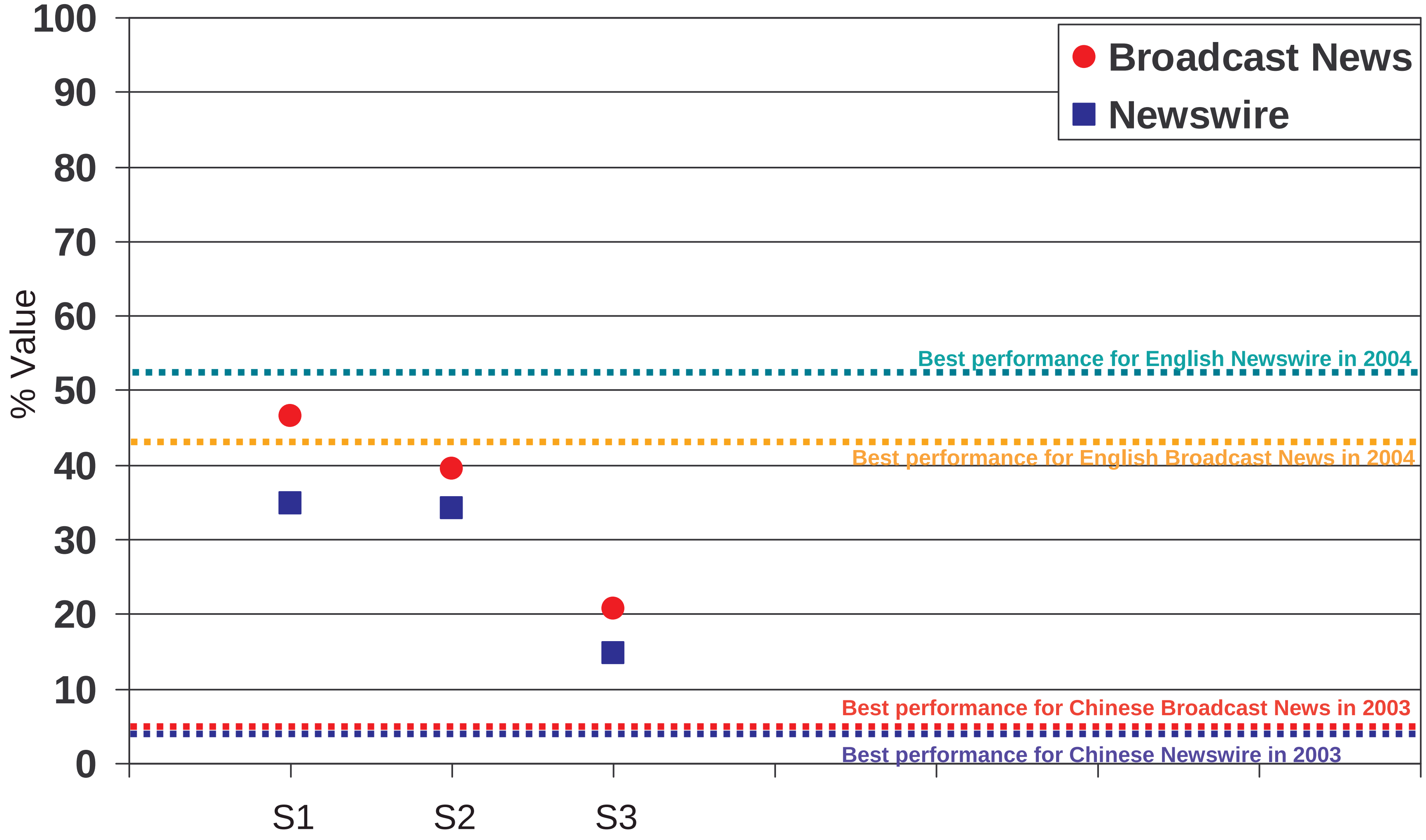
2004 RDR Results English – Source Set - systems ordered by decreasing average % value (combined)

RDR Results for *Arabic*



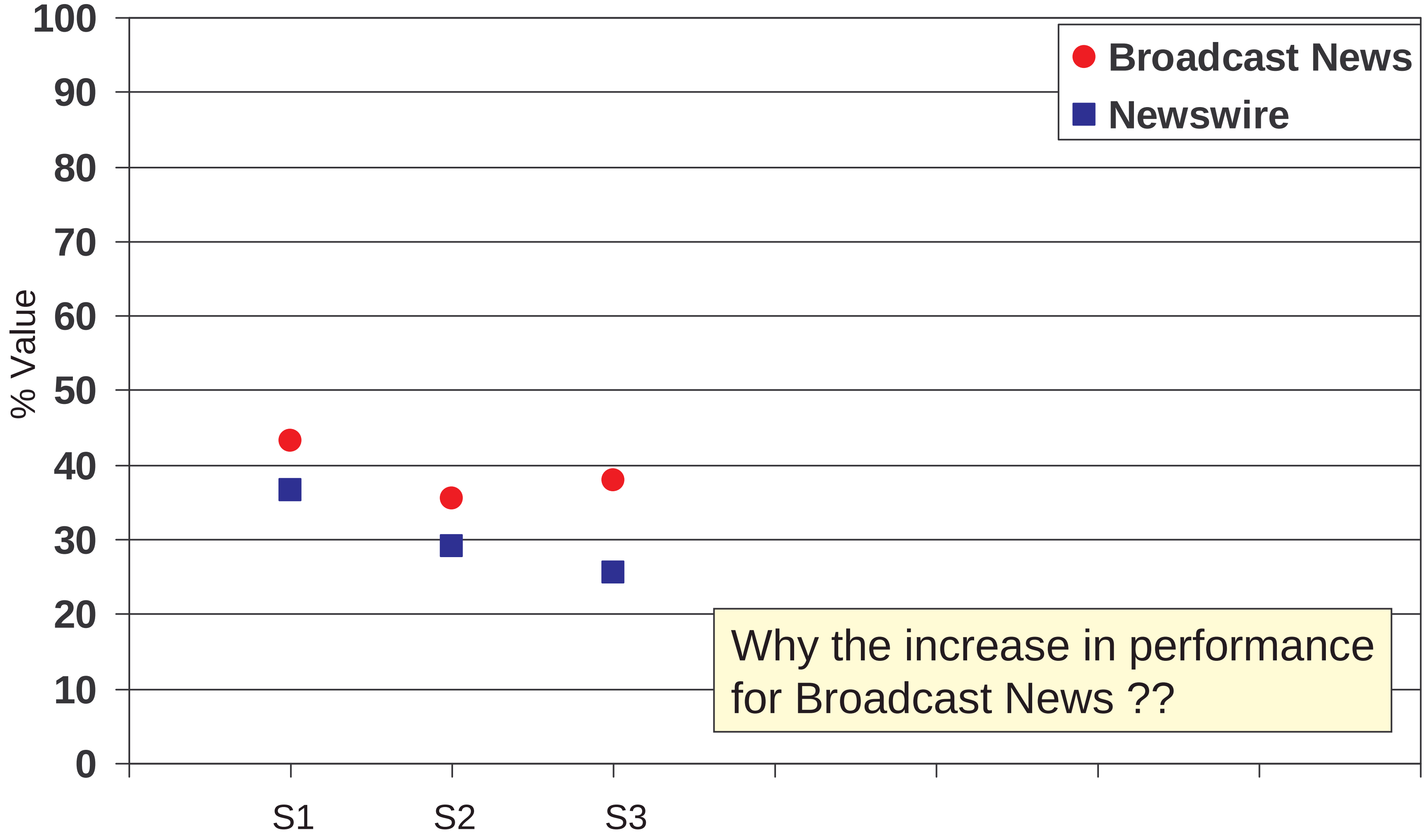
2004 RDR Results Arabic – Source Set - systems ordered by decreasing average % value (combined)

RDR Results for *Chinese*



2004 RDR Results Chinese – Source Set - systems ordered by decreasing average % value (combined)

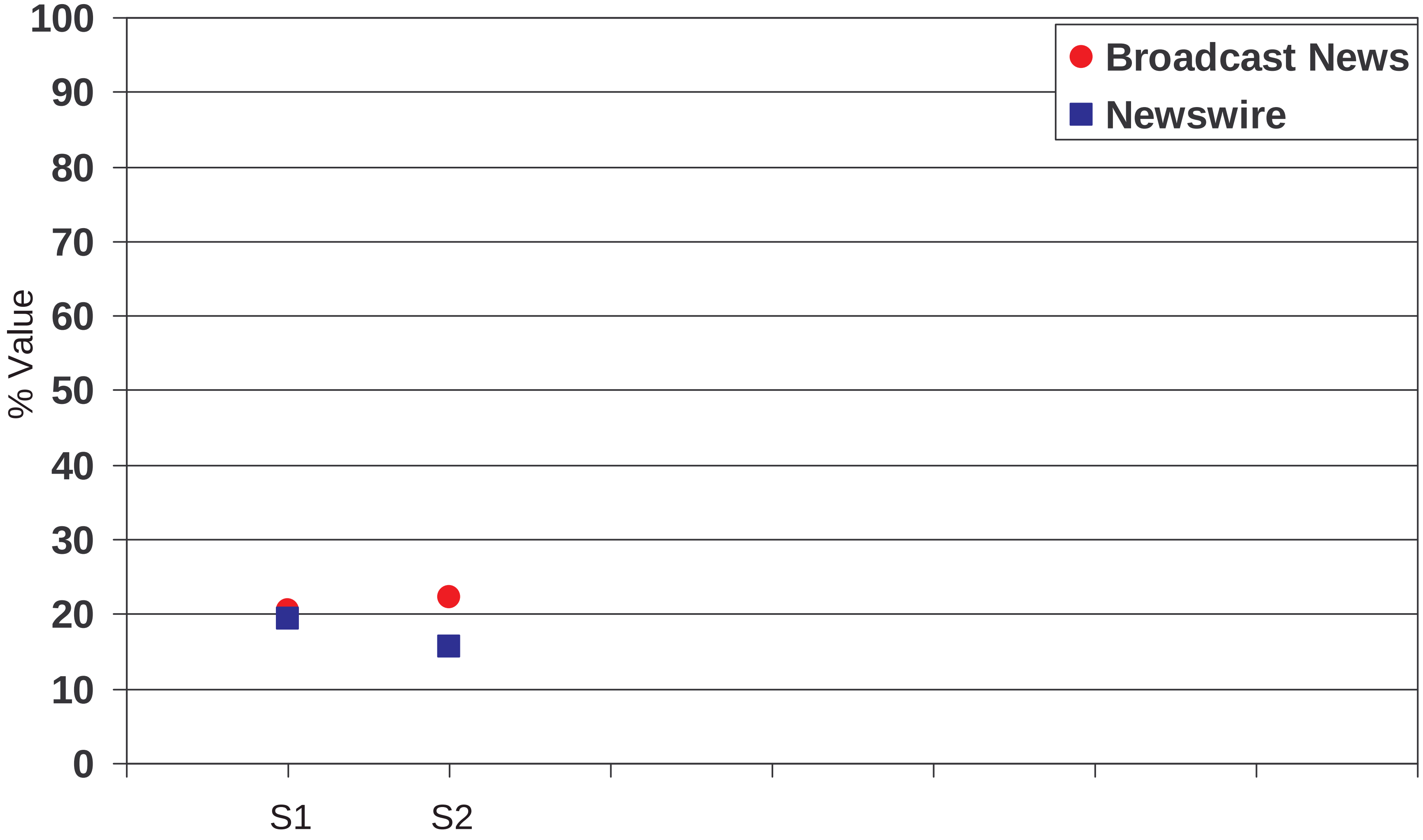
RMD Results for *English*



Why the increase in performance for Broadcast News ??

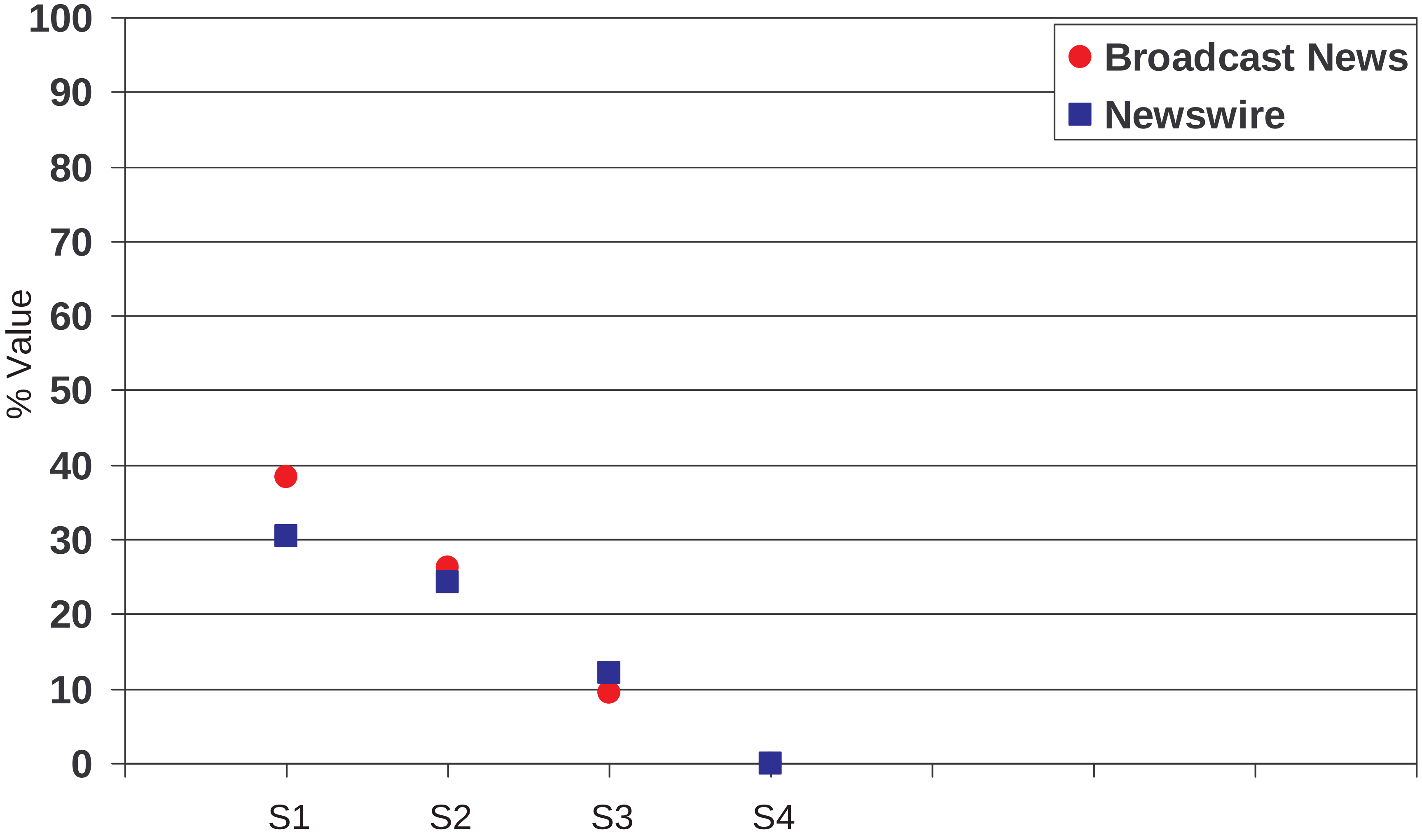
2004 RMD Results English – Source Set - systems ordered by decreasing average % value (combined)

RMD Results for *Arabic*



2004 RMD Results Arabic – Source Set - systems ordered by decreasing average % value (combined)

RMD Results for *Chinese*



2004 RMD Results Chinese – Source Set - systems ordered by decreasing average % value (combined)

EDR

Analysis by ***entity type***

for each of

the

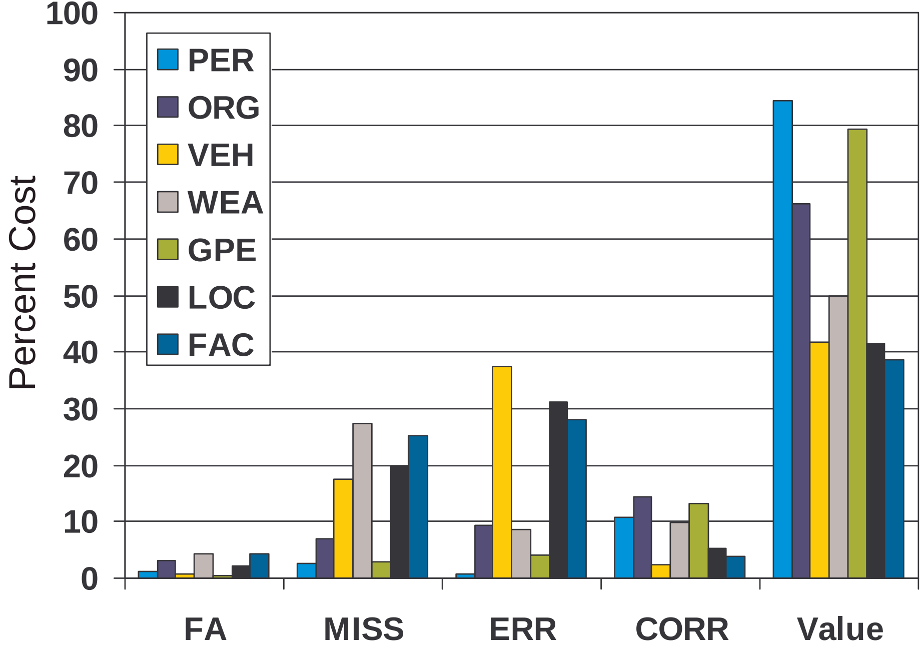
three languages

English EDR

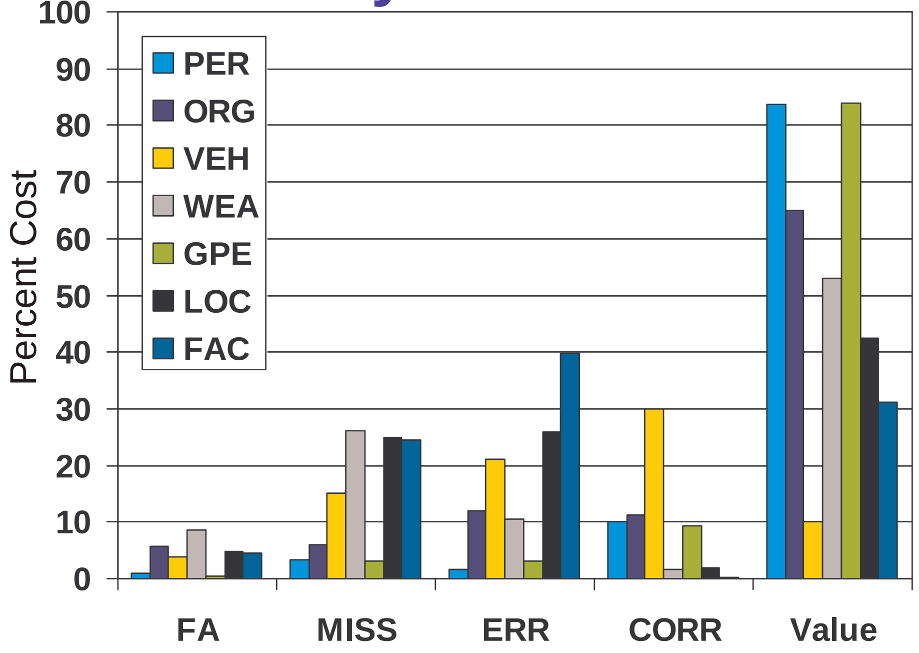
Percent of Cost by Type

Combined Sources (Broadcast News & Newswire)

System ?



System ?



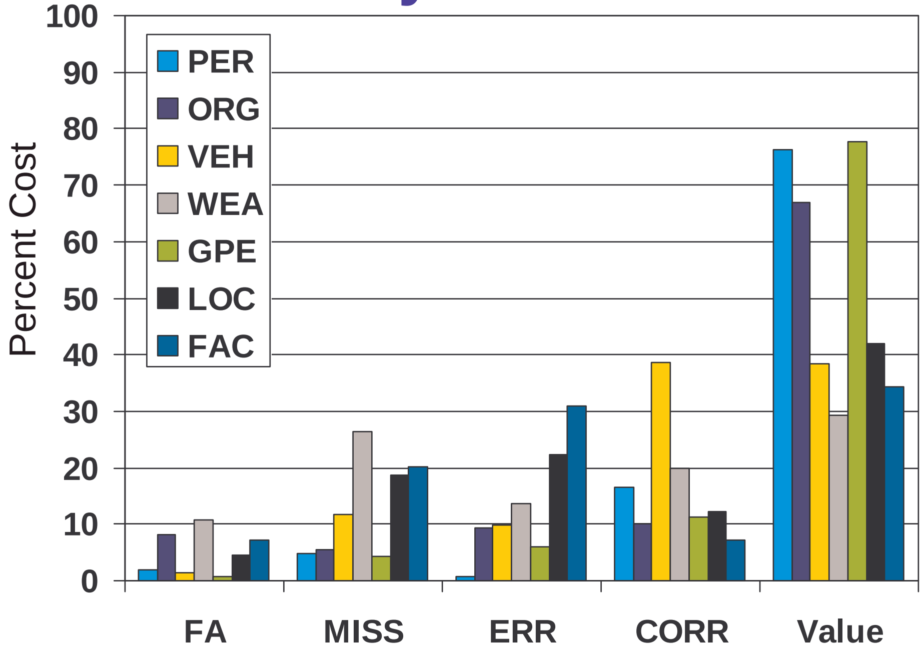
- Non-zero values in the “corr” column are due to missing or spurious mentions in the system output entity

Arabic EDR

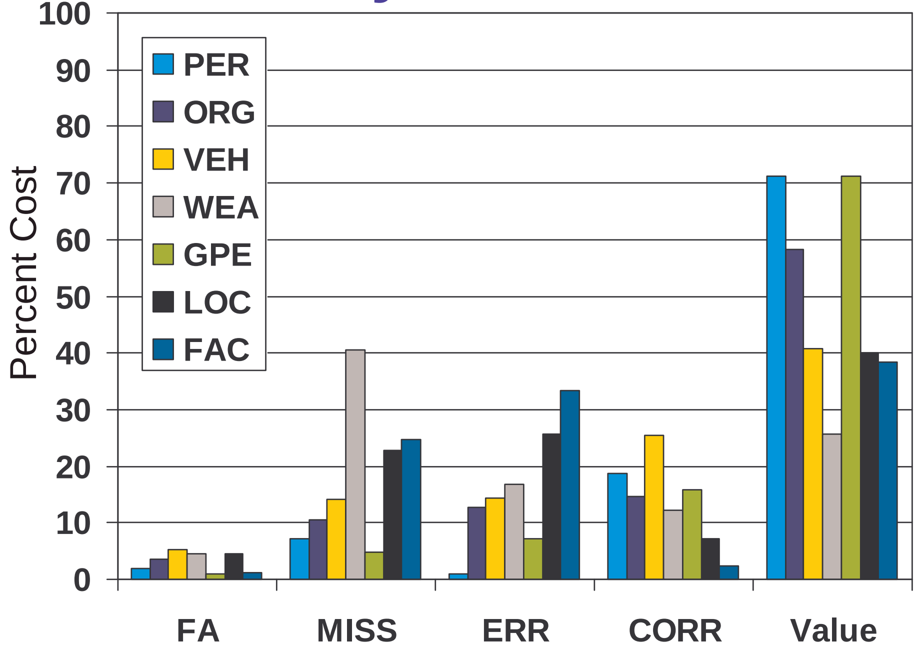
Percent of Cost by Type

Combined Sources (Broadcast News & Newswire)

System ?



System ?

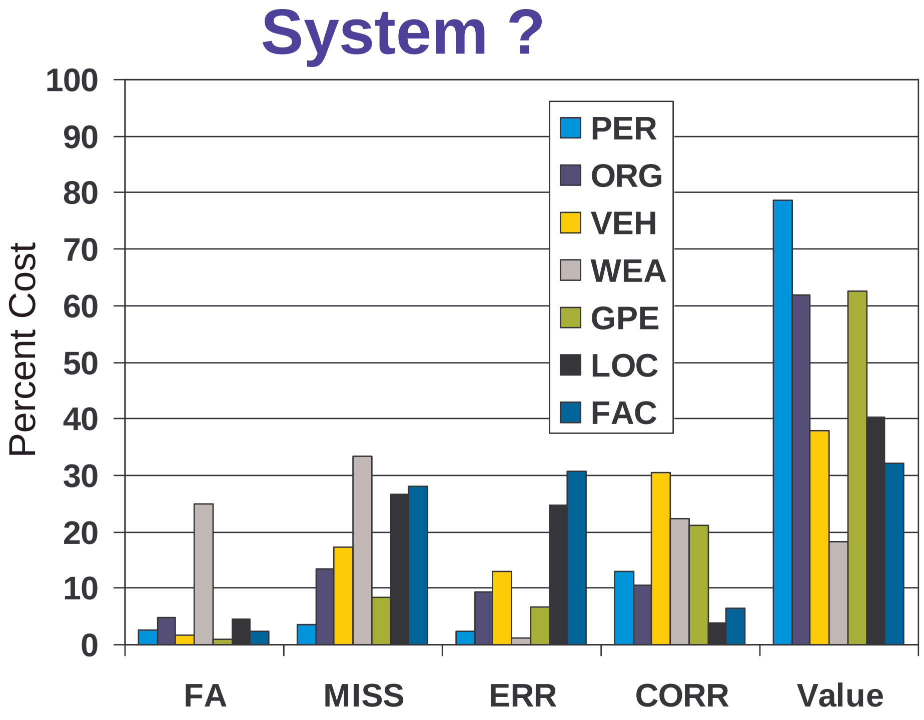


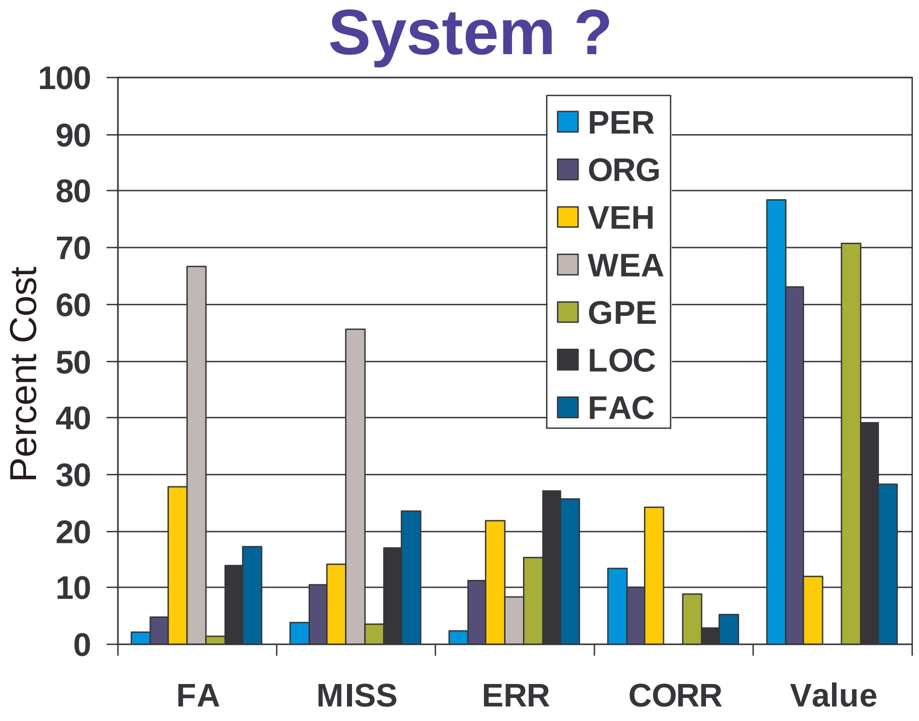
- Similar error distribution among top systems

Chinese EDR

Percent of Cost by Type

Combined Sources (Broadcast News & Newswire)





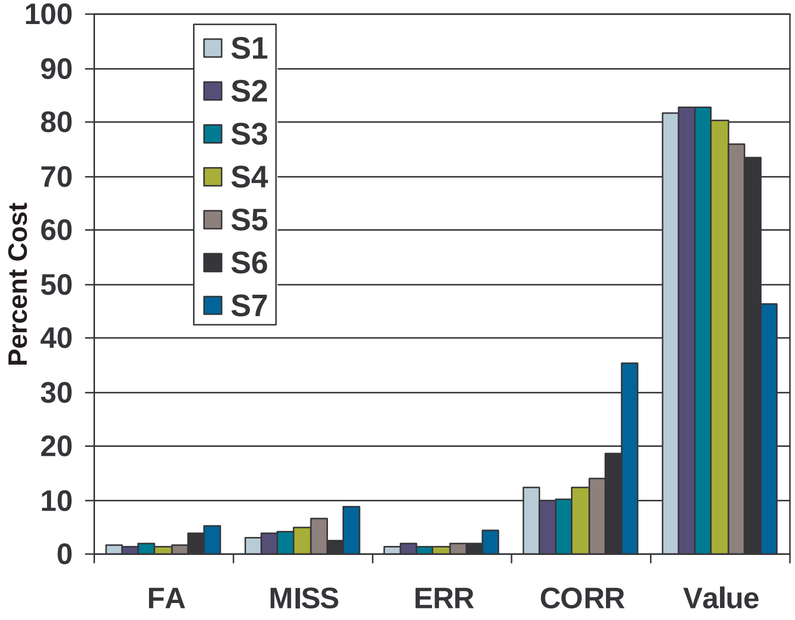
- Similar error distributions across language (compare this slide with the previous two)

EDR – *English*

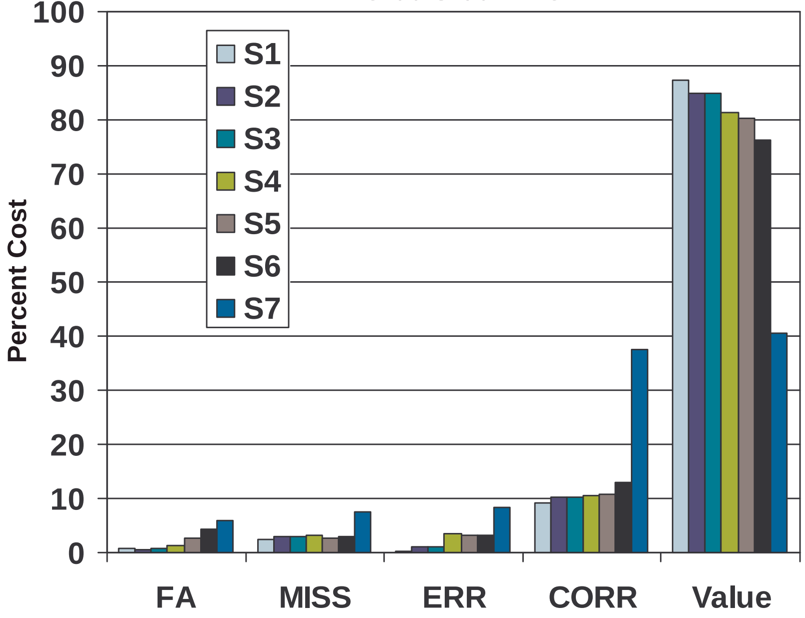
Entity Type - *Person*

PER Value (%)	S1	S2	S3	S4	S5	S6	S7
BNEWS (1222)	81.6	***82.8***	***82.8***	80.2	75.8	73.3	46.3
NWIRE (865)	***87.4***	84.9	84.8	81.3	80.4	76.3	40.5

Broadcast News



Newswire

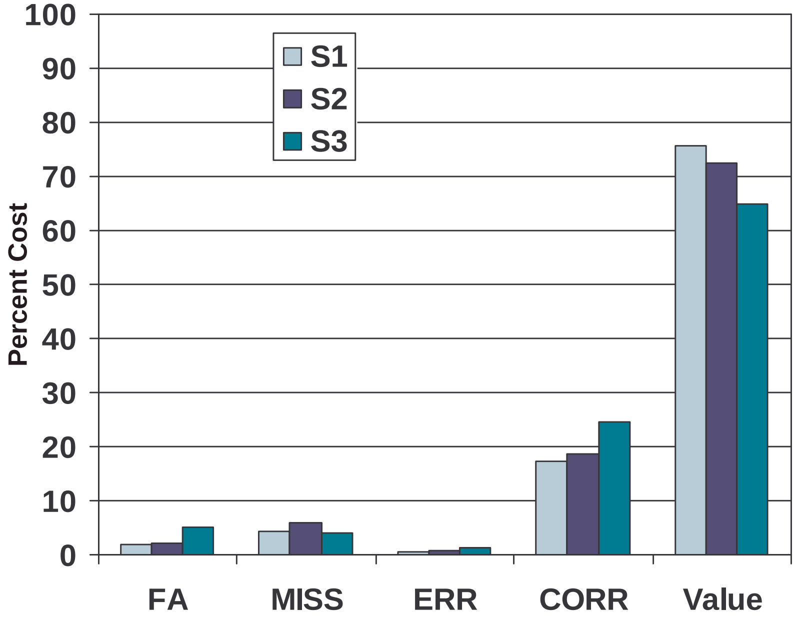


EDR – *Arabic*

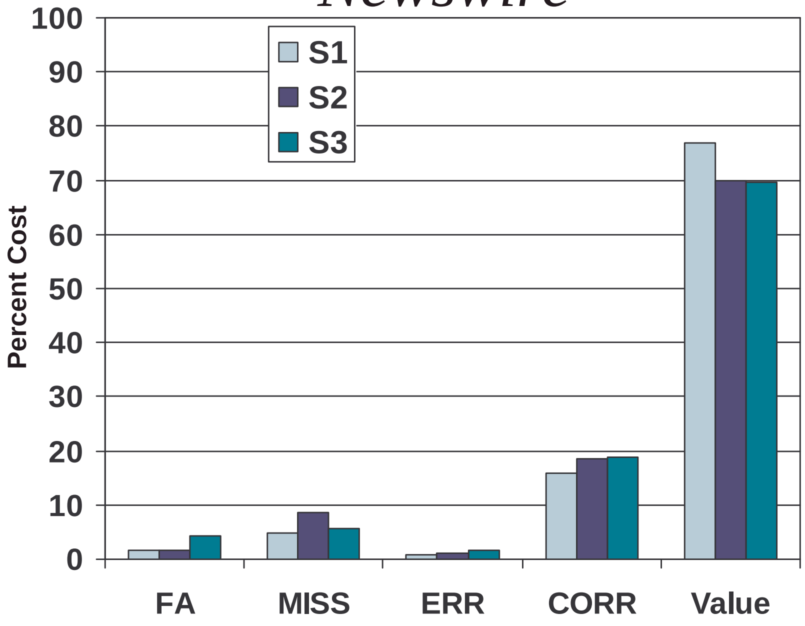
Entity Type - *Person*

PER Value (%)	S1	S2	S3
BNEWS (1245)	***75.7***	72.4	64.9
NWIRE (1217)	***76.8***	70.0	69.6

Broadcast News



Newswire

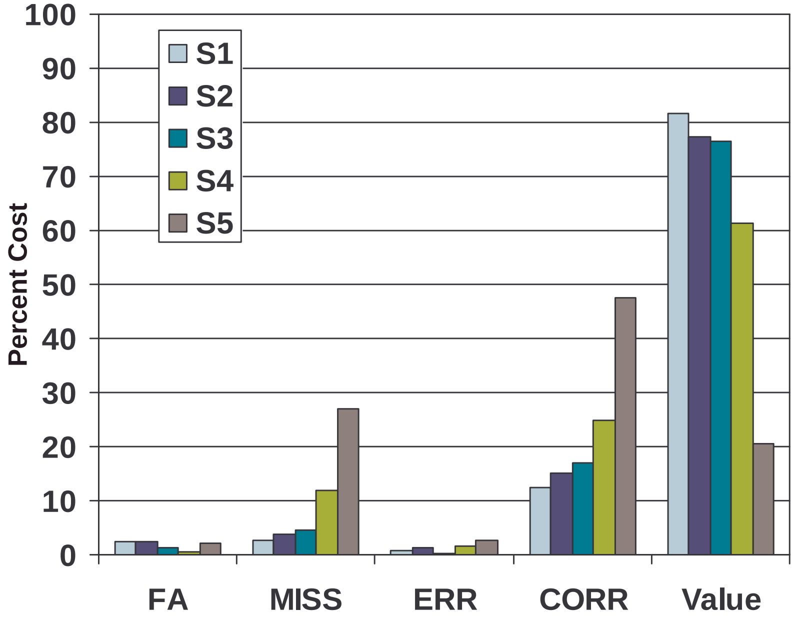


EDR – *Chinese*

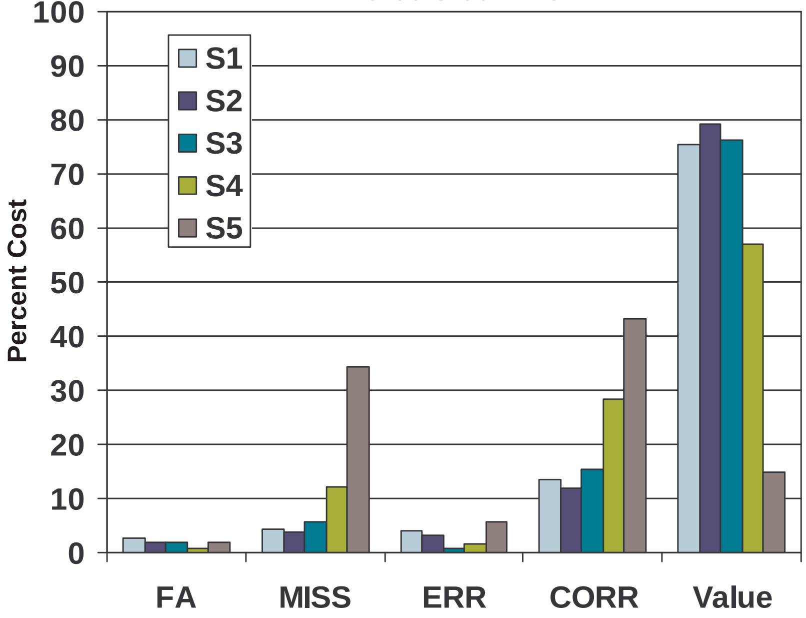
Entity Type - *Person*

PER Value (%)	S1	S2	S3	S4	S5
BNEWS (952)	***81.6***	77.4	76.6	61.3	20.6
NWIRE (1030)	75.5	***79.3***	76.2	57.1	15.0

Broadcast News



Newswire



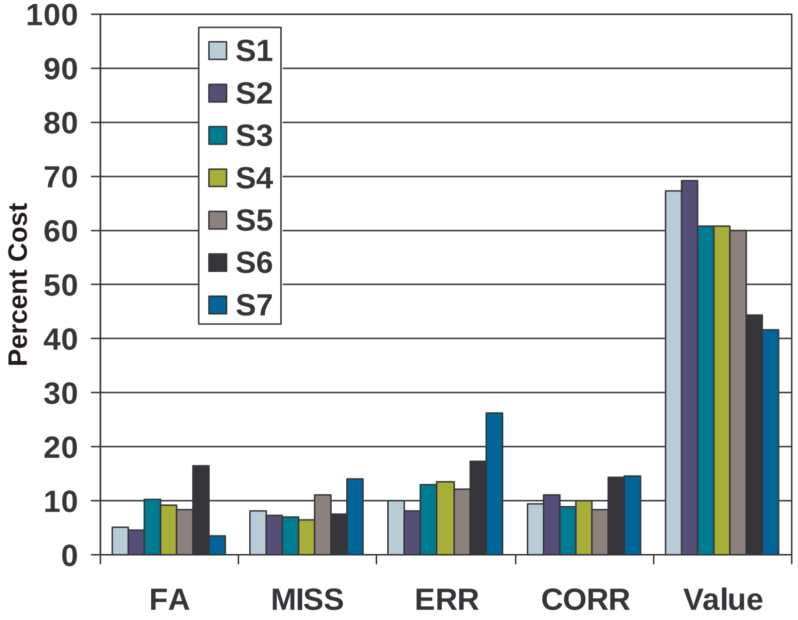
EDR – *English*

Entity Type - *Organization*

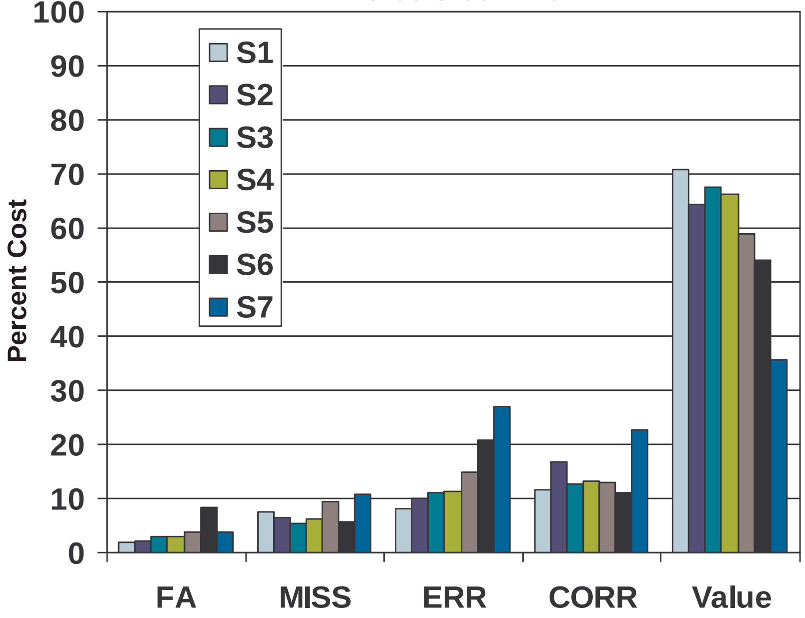
ORG Value (%)	S1	S2	S3	S4	S5	S6	S7
BNEWS (394)	*67.2*	***69.1***	60.9	60.0	59.9	44.2	41.5
NWIRE (518)	***70.7***	64.3	67.7	66.3	58.9	54.1	35.8

- Mid-80% of value for PERSON down to about 70% for ORGANIZATION

Broadcast News



Newswire

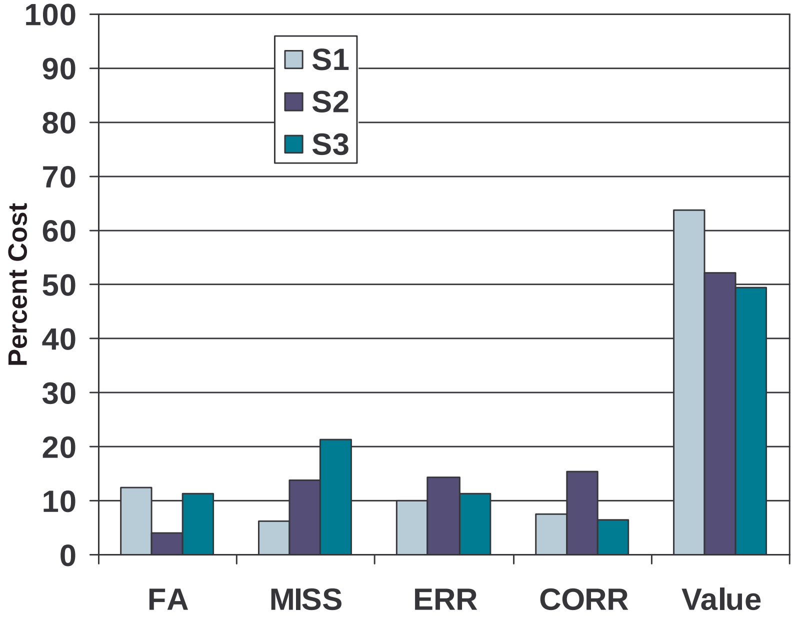


EDR – *Arabic*

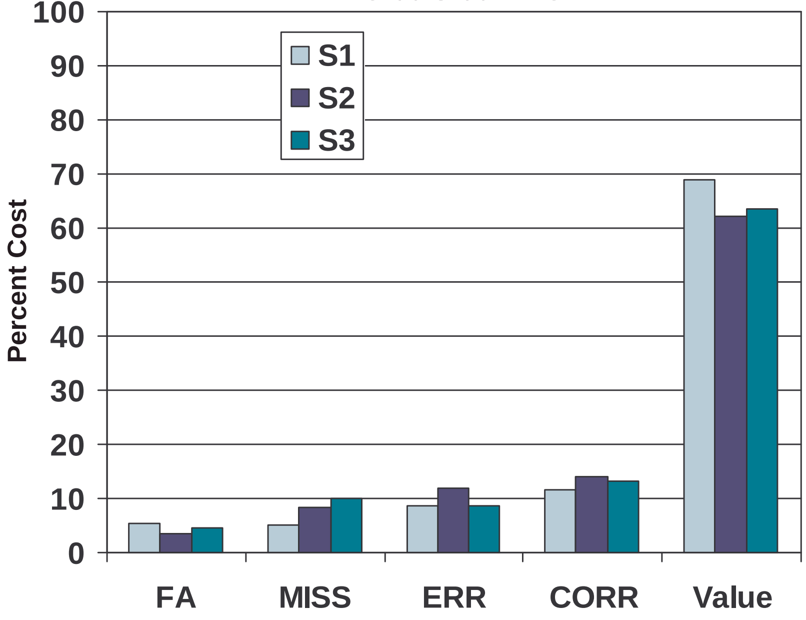
Entity Type - *Organization*

PER Value (%)	S1	S2	S3
BNEWS (441)	***63.7***	52.2	49.5
NWIRE (558)	***68.8***	62.2	63.4

Broadcast News



Newswire

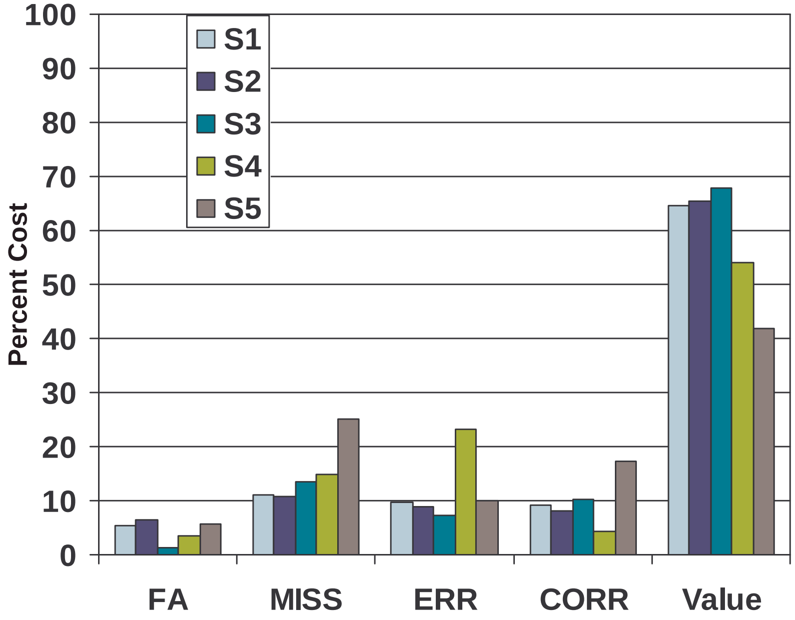


EDR – *Chinese*

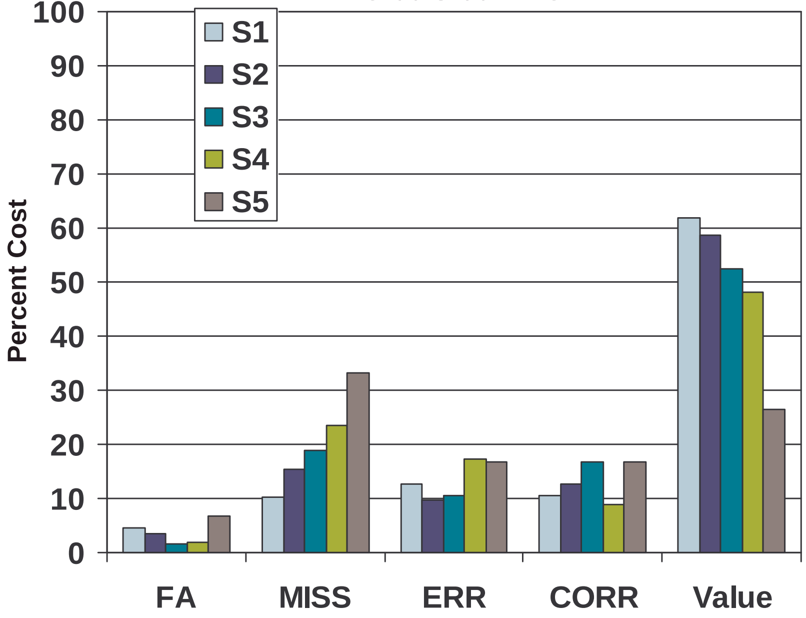
Entity Type - *Organization*

PER Value (%)	S1	S2	S3	S4	S5
BNEWS (593)	***64.6***	65.4	67.8	54.0	41.9
NWIRE (685)	***61.9***	58.7	52.3	48.1	26.4

Broadcast News



Newswire



Summary

- Lots of data results to talk about
  - Some analysis in hand out only
- PER, ORG, GPE major contributors to overall value
- Exercise caution when trying to draw conclusions on progress
  - different scorers, and
  - changes in the task definition.
- 24 hour turn around on results worked well, maybe we don’t need the two week window?
- Did not cover Diagnostic Tasks
  - EDR Co-reference (given ground truth mentions)
  - RDR given ground truth entities